INVESTIGATIONS IN PHYLOGENETICS: TREE INFERENCE AND MODEL IDENTIFIABILITY

By

Samaneh Yourdkhani

A Dissertation Submitted in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Mathematics

University of Alaska Fairbanks

May 2020

APPROVED:

Dr. John A. Rhodes, Committee Co-Chair
Dr. Elizabeth S. Allman, Committee Co-Chair
Dr. Julie McIntyre, Committee Member
Dr. Gordon Williams, Committee Member
Dr. Leah Berman, Chair
Department of Mathematics and Statistics
Dr. Kinchel C. Doerner, Dean
College of Natural Science and Mathematics
Dr. Michael Castellini, *Dean of the Graduate School*

Abstract

This thesis presents two projects in mathematical phylogenetics. The first presents a new, statistically consistent, fast method for inferring species trees from topological gene trees under the multispecies coalescent model. The algorithm of this method takes a collection of unrooted topological gene trees, computes a novel intertaxon distance from them, and outputs a metric species tree. The second establishes that numerical and non-numerical parameters of a specific Profile Mixture Model of protein sequence evolution are generically identifiable. Algebraic techniques are used, especially a theorem of Kruskal on tensor decomposition.

Contents

List of Figures

Acknowledgments

I'm extremely grateful to the completion of my dissertation which would not have been possible without the support of my great advisors, Dr. John Rhodes and Dr. Elizabeth Allman. They provided me with encouragement and patience throughout the duration of my PhD project.

I also would like to thank my committee members, Dr. Gordon Williams, Dr. Julie McIntyre and the chair of the Department of Mathematics and Statistics, Dr. Leah Berman for taking the time to review and give suggestions for this work.

At Last, many thanks to my family who always supported and nurtured me.

Chapter 1: Introduction

Phylogenetics is concerned with the evolutionary relationships among species. In mathematical phylogenetics, we use mathematical tools to solve problems related to the reconstruction and analysis of phylogenetic trees. This work addresses some challenges in phylogenetics using a range of mathematics tools from probability, combinatorics, and geometry. In this work two different aspects of mathematical phylogenetics are studied.

For the first project,"Fast inference of metric species trees from topological gene trees," the relationships between gene trees relating biological sequences and species trees relating populations is investigated, since it focuses on improving methods for reconstruction of species trees from gene trees under a specific model.

When we have DNA, or other sequences, for one gene in individuals from different species, a gene tree can be inferred that shows the evolutionary relationships of these individual sequences. A species tree shows the evolutionary relationship between the species overall, but because of biological processes such as incomplete lineage sorting, gene trees do not always match the species trees. This is illustrated in Figure 1.1.

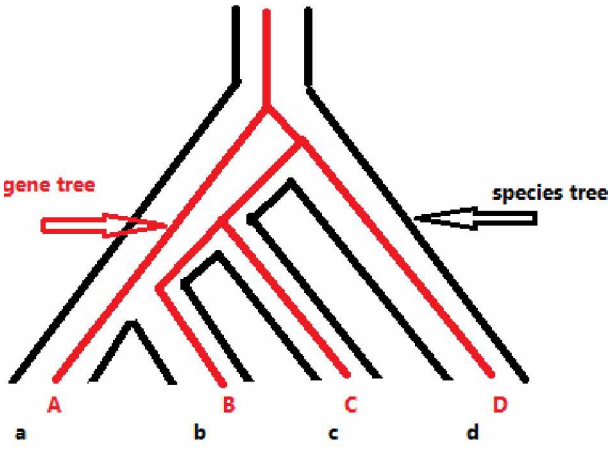

Figure 1.1: A gene tree relating lineages A, B, C, and D within the species tree.

The conflict between gene trees and species trees makes the inference problem for species relationships challenging as standard phylogenetic methods applied to different genes for the

same collection of species often give different trees. The highlight of our finding is even with only topological information on gene trees and no information about the species tree, a fast combinatorial method is guaranteed to produce the correct metric species tree under ideal circumstances. After developing the theory, we analyzed simulated data using `R` to see how the ideas might work in practice, and got good results.

The second project addresses whether it is theoretically possible to recover the parameters of a particular probabilistic model used to infer phylogenetic trees from protein sequence data if given an infinite number of observations. The model in this project, called the "Profile Mixture Model" has a huge number of parameters, since some biologists feel this complexity is needed to capture the essential features of actual evolutionary changes [1]. For an n-taxon tree, the number of numerical parameters is more than $1448 + 2n - 3$, which makes this problem very complicated.

The techniques used in this project come from a wide range of topics in computational algebra, algebraic geometry, and graph theory. These include Kruskal rank, algebraic varieties, tensor products, continuous-time Markov processes, and some computer software for exact computations.

Chapter 2 of this thesis gives a detailed explanation of the multispecies coalescent model and contains the paper "Fast inference of metric species trees from topological gene trees". Chapter 3 explains the Profile Mixture Model identifiability problem and contains the results of the second project. Chapter 4 addresses some future questions.

Chapter 2: Inferring Species Tree from Gene Trees[1]

As mentioned in Chapter 1, a gene tree often differs from the species tree, due to a population genetic phenomenon called *Incomplete Lineage Sorting* (ILS). The model that gives the framework for inferring species trees from gene trees accounting for such gene tree-species tree conflict is the *Multispecies Coalescent Model, MSC.*

To describe this model, we first talk about a simpler discrete one that is the *Wright-Fisher* model. In this model, we consider N individuals in a single population at all times, where time corresponds to generations, as shown in Figure 2.1. Each row represents a generation, and each dot, a gene in the population in that generation.

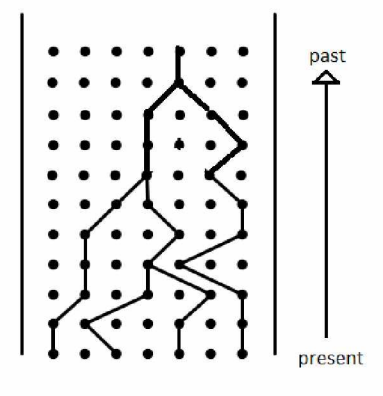

Figure 2.1: A Wright-Fisher simulation: N of individuals in a single population at all time.

We move backwards in time (from the present to the past) and suppose that each gene picks a parent uniformly at random from the previous generation. It is easy to find the probability that two lineages coalescence in the parental generation since regardless of which parent the first gene picks, there is just one choice out of N for the second gene that gives coalescence. Then the probability that they will coalesce immediately in the previous generation is $1/N$. By the same reasoning, we can compute the probability that two specific extant lineages coalescence in generation n that is $\left(1-\frac{1}{N}\right)^{n-1}\frac{1}{N}$. Then the expected number of generations for coalescing of two specific lineages is N, which means when the number

[1]Part of this chapter has been submitted to the Bulletin of Mathematical Biology as a joint publication with J. A. Rhodes.

of individuals in the population gets larger, it takes longer to have a coalescent event on average. But since the probability of coalescence of two specific lineages by generation n is $1 - \left(1 - \frac{1}{N}\right)^n$, they will eventually coalesce at some time because as n goes to infinity, this probability approaches 1.

Since the expected number of generations for two lineages to coalesce is N, the population size, a small number of generations in a small population has the same impact on coalescence as a larger number of generations in a large population. This motivates a new time scale for each population that is the number of generations divided by the population size. This time scale in the continuous model is saved to be in *coalescent units*. Now using continuous time instead of discrete time, the Wright-Fisher model gives us *Kingman*'s coalescent model and the extension of that to a species tree of populations is called the MSC model.

In the Kingman or MSC model with time measured in coalescent units, the instantaneous rate of coalescence is 1. We also suppose that no more than two lineages may coalesce simultaneously and coalescence of different pairs is independent, and identically distributed. Since the coalescent events occur rarely we assume they occur as a Poisson process. Let $h(u)$ be the probability that two distinct lineages do not coalesce between time 0 and time $u > 0$. From basic facts about Poisson processes and the fact that no coalescent event happens at time 0, we have $h(u) = e^{-u}$. Since there are $\binom{n}{2}$ pairs of n extant lineages, then the probability that all n lineages remain distinct at time u is

$$h_n(u) = e^{-\binom{n}{2}u}.$$

Now we are able to compute the probabilities of observing different topological gene trees. Although it is complicated to compute such probabilities for large trees, we begin with a 3-sample tree that is a crucial feature in the next part. As it can be seen in Figure 2.2, there are 3 possible gene trees that may relate samples of one gene per taxon from a 3-taxon species tree.

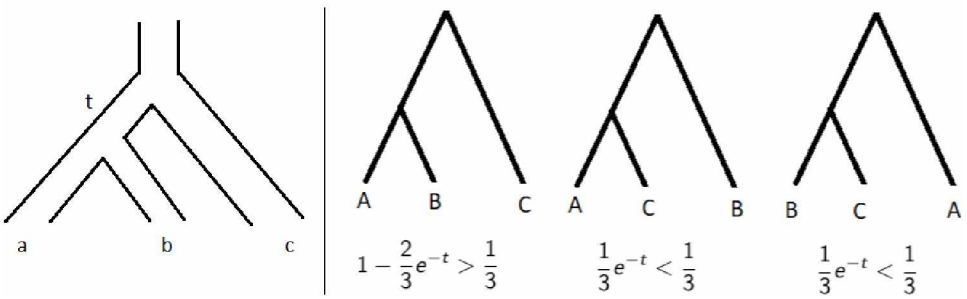

Figure 2.2: A 3-taxon species tree $((a,b),c)$ with 3 possible topological gene trees with one gene sampled per taxon, A, B, C.

Since no coalescent event can occur in the terminal edges on the species tree with any one sampled lineage in them, then in the probability formulas these edges do not appear. Also we know that the probability of two lineages A, B not coalescing in t coalescent units is e^{-t}, and if the A, B, C lineages all reach the root population on the species tree the probability of A, B coalescing first is $1/3$. Thus, the probability of observing either of the two discordant gene trees is $\frac{1}{3}e^{-t}$. Since the probabilities of observing these 3 topological gene trees must add to 1, then the matched gene tree has the probability $1-\frac{2}{3}e^{-t}$. It has been proved [**?**] that for 4-taxon unrooted trees, the same probability formulas for the 3 possible unrooted topological gene trees are satisfied.

We next present the paper "Inferring Metric Trees From Weighted Quartets via an Intertaxon Distance." This paper introduces ways of remetrizing a metric tree based on its 4-taxon subtrees, which is primarily a combinatorial result. However the motivation and the application of this remetrization is to inferring a species tree from gene trees under the MSC model.

INFERRING METRIC TREES FROM WEIGHTED QUARTETS VIA AN INTERTAXON DISTANCE

SAMANEH YOURDKHANI AND JOHN A. RHODES

Abstract. A metric phylogenetic tree relating a collection of taxa induces weighted rooted triples and weighted quartets for all subsets of three and four taxa, respectively. New intertaxon distances are defined that can be calculated from these weights, and shown to exactly fit the same tree topology, but with edge weights rescaled by certain factors dependent on the associated split size. These distances are analogs for metric trees of similar ones recently introduced for topological trees that are based on induced unweighted rooted triples and quartets. The distances introduced here lead to new statistically consistent methods of inferring a metric species tree from a collection of topological gene trees generated under the multispecies coalescent model of incomplete lineage sorting. Simulations provide insight into their potential.

1. Introduction

We introduce new intertaxon distances that are computed for taxa on an unrooted metric phylogenetic tree based on its displayed rooted triples or quartets. The distances depend upon the *weights* — the lengths of the unique internal edge — of the rooted triples or quartets. These distances differ from the original intertaxon distance on the metric tree, but exactly fit the tree topology, allowing standard distance methods to be used to recover the tree from knowledge of only its weighted rooted triples or quartets. If the rooted triple or quartet data is noisy, so that not all are correct, this distance can still be used to estimate the tree. While the tree estimate will have edge lengths estimating those on a remetrized tree, a simple adjustment gives estimates of the original edge lengths. Thus these distances lead to new distance-based consensus methods for obtaining a large metric tree from a collection of weighted rooted triples or quartets. In particular they can be used in new statistically consistent methods of metric species tree inference from topological gene trees under the standard model of incomplete lineage sorting.

This final application is, in fact, our motivation for developing these distances. Statistical inference of a species tree under the multispecies coalescent (MSC) model of incomplete lineage sorting is a fundamental problem in current phylogenetic data analysis. For large datasets (many taxa, with sequences from many loci) that are increasingly common in empirical studies, the simultaneous inference of gene and species trees by Bayesian methods [Liu08, HD10] may require excessive computation time. Other methods proceed by first inferring gene trees for each locus, and then treating these as data for a second inference of the species tree [VW15, ZRSM18].

Date: February 6, 2020.

This work continues a thread of developments initiated with several methods of this second sort introduced by Liu and collaborators [LYPE09, LY11] for inferring a species tree from a collection of topological gene trees, either rooted or unrooted, under the MSC model. These methods, called STAR and NJ$_{st}$, proceed by first remetrizing the gene trees in a way that reflects only their topologies, next computing intertaxon distance matrices from each remetrized tree, and then averaging these matrices. Finally, a standard distance method such as Neighbor Joining is used to construct a species tree from this average distance. Despite this seemingly simplistic approach, the methods are statistically consistent under the MSC model [ADR13, ADR18], and show strong performance in simulation studies [VW15]. Moreover, they have been shown to be based on the underlying notions of displayed clades and splits on the gene trees [ADR13, ADR18]. A third method, STEAC [LYPE09], took a similar averaging approach while retaining metric information on the gene trees. Its statistical consistency, however, requires assumptions on the relationship of gene tree metric units (substitution units) to species tree metric units (coalescent units) which may be difficult to justify.

Motivated by the STAR and NJ$_{st}$ algorithms, the RTDC and QDC methods [Rho19] are based on similar distances defined from displayed topological rooted triples and quartets on gene trees, and give statistically consistent inference of topological species trees from gene trees under the MSC. Although the use of the quartet and rooted triple distances result in a slower algorithm than the split or clade approaches of STAR and NJ$_{st}$, inference with them is more robust to missing taxa on gene trees, and gives similar performance to, for instance, the highly developed quartet-based inference software ASTRAL. Moreover, the quartet distance has been generalized to the level-1 network setting [ABR19], playing a key role in the NANUQ method for fast inference of hybridization networks.

While the results presented here are analogs for metric trees of the results for topological trees of [Rho19], the remetrizations we develop are genuinely new, and not simple extensions of the topological quartet and rooted triple ones. Moreover, since the weights in coalescent units of rooted triples and quartets can be inferred from *topological* gene tree data under the MSC, one can estimate these new intertaxon distances on a species tree from topological gene trees alone. Thus from the same gene tree data considered in [Rho19], one obtains not only an estimate of the topology of the species tree, but a metric estimate as well. While the ability to infer a metric species tree is thus similar to STEAC's, the approach introduced here crucially uses no *metric* gene tree information, and thus its consistency does not depend on any assumptions of the relationship of metric units on gene trees and the species tree. It is thus statistically consistent under much broader assumptions. Although the limited simulation results we present indicate that further work will be necessary to produce algorithms competitive with other approaches, these distances provide new tools for understanding how information on a species tree can be extracted from the gene trees.

Although we position this work in the context of species tree inference, the basic problem of inferring a tree from weighted quartets is not new. Characterizations of those weighted quartet systems that define a metric tree have been given for both binary [DE03] and and non-binary [GHMS08] trees, in settings where all weights are known exactly. The

weighted quartet distance defined here offers advantages in any setting where there may be noise in the weights, and an exact fit to a single tree is not possible. Then any of the many methods of fitting a tree to a distance matrix may be applied for an approximate solution.

The remainder of this paper proceeds as follows. After introducing notation and definitions in Section 2, the weighted rooted triple metrization and its associated distance is developed in Section 3. Section 4 develops the analogs for weighted quartets. Several algorithms using these distances for the inference of a tree from its displayed quartets or a collection of gene trees are formalized in Section 5. Finally, Section 6 presents some preliminary simulation results, and discusses some of the practical issues of using these distance for inference.

Implementations of the quartet versions of the algorithms developed and used in this paper are available in the R package `MSCquartets` [ABMR19].

2. Background and Notation

By a *rooted topological phylogenetic tree* T^r on X we mean a rooted tree whose root has degree ≥ 2 and all other internal nodes have degree ≥ 3, with leaves bijectively labelled by elements of the finite taxon set X. Directing edges away from the root, we have an ancestral partial order on the nodes, with the root ancestral to all others.

A *rooted metric phylogenetic tree* (T^r, λ^r) on X is a rooted topological tree together with a function λ^r which assigns non-negative weights, or *edge lengths*, to all edges of T^r. We use T and (T, λ) to denote the unrooted topological and metric species trees obtained from T^r and (T^r, λ^r) in the obvious way, by suppressing the root node if it has degree 2, and undirecting edges.

The *most recent common ancestor* of taxa $x, y \in X$ on a rooted tree T^r is a the minimal node ancestral to both, denoted $\mathrm{MRCA}(x, y)$. By the *descendants* of a node v, denoted $\mathrm{desc}(v)$, we mean the subset of X labelling leaves that have v as an ancestor.

When considering the *multispecies coalescent model* (MSC) [PN88], we denote its species tree parameter by (σ^r, λ^r). Edge lengths on a species tree are measured in *coalescent units*, which are units of time (in generations) inversely scaled by population size, so that the rate of coalescence of two gene lineages in an edge (i.e., population) on the species tree is normalized to 1. Such a parameter determines a probability distribution on rooted and unrooted topological gene trees on X, which we denote as T^r or T. Under the MSC non-binary topological gene trees have probability 0 even when the species tree is non-binary. Assuming one gene lineage is sampled for each taxon in X, the topological tree σ^r and the edge lengths $\lambda^r(e)$ for all internal edges e on σ^r are identifiable from the distribution of rooted topological gene trees T^r, although lengths of pendant edges on σ^r are not. In fact, σ^r and $\lambda^r(e)$ are identifiable for internal edges e even from the distribution of unrooted topological gene trees T when $|X| \geq 5$. However, if $|X| = 4$ only the unrooted σ and its one internal edge length are identifiable [ADR11].

A resolved *rooted triple* is a 3-taxon rooted tree, denoted by $ab|c = ba|c$ where the taxa a, b form a clade. The unresolved rooted triple, a star tree on a, b, c is denoted abc. A

rooted tree σ^r or T^r on X *displays* the rooted triples it induces on 3-taxon subsets of X. A *weighted rooted triple* is a pair of a rooted triple together with a weight, a non-negative real number. We view the weight for a resolved rooted triple as a length for the single internal edge of the triple, and allow a weight of zero only if the rooted triple is unresolved. A rooted triple $ab|c$ is said to *separate* the pair a and c, as well as the pair a and b. An unresolved rooted triple does not separate any pairs of taxa on it. The set of rooted triples on X separating taxa a, b is denoted $\mathcal{RT}_{ab}$, and the subset of these rooted triples displayed on T^r by $\mathcal{RT}_{ab}(T^r)$.

Similarly, a resolved *quartet* is a 4-taxon unrooted tree, denoted by $ab|cd = ba|cd = ab|dc = ba|dc$ where the taxa a, b and c, d form cherries. The unresolved quartet, a star tree on a, b, c, d is denoted $abcd$. An unrooted tree σ or T *displays* the quartets it induces on 4-taxon subsets. A *weighted quartet* is a pair of a quartet together with a weight, a non-negative real number. We view the weight for a resolved quartet as a length for the single internal edge of the quartet tree, and only allow the weight 0 for the unresolved quartet. A quartet $ab|cd$ is said to *separate* the taxon pair a and c, as well as the pairs a, d and b, c and b, d. An unresolved quartet does not separate any pairs of taxa on it. The set of quartets on X separating taxa a, b is denoted $\mathcal{Q}_{ab}$, and the subset of these quartets displayed on T by $\mathcal{Q}_{ab}(T)$.

Any metric tree (T^r, λ^r) or (T, λ) on X induces a metric d_λ on X, using the sum of edge weights along paths between the taxa. As is well known, however, a metric d on X need not arise from such a weighting. If $d = d_\lambda$ for some λ on T, then we say d is a *tree metric* on T with weighting λ.

For nodes v and w on T, define $P_{v,w} = \{e_1, e_2, \ldots, e_k\}$ to be the path from v to w on T. For a rooted tree T^r, we use the same notation for the set of edges which forms a path from v to w when undirected.

3. Weighted rooted triple metrization of a rooted tree

Given a rooted metric tree, we introduce a remetrization of the tree, so that internal edge lengths become a product of their original lengths and an integer factor dependent on the placement of the edge in the topological tree. Although this introduces no new information, the value of doing this, which will be developed in later section, is to enable an algorithmic approach to inferring a metric tree from its weighted rooted triples, even in the presence of noise. The key theoretical underpinning of this is Theorem 3.1 of this section.

Let (T^r, λ^r) be a rooted metric phylogenetic tree on X. For any vertex v on T^r, denote by $n(v)$ the number of taxa in X which are *not* descendants of v. We remetrize T^r to obtain a new metric tree $\left(T^r, \tilde{\lambda}^r\right)$ as follows: First for each internal edge $e = (u, v)$ with u the parent of v let

$$\tilde{\lambda}^r(e) = \lambda^r(e) \cdot n(v). \tag{1}$$

Then assign pendant edge lengths in such a way that the tree becomes ultrametric (i.e. all root-to-leaf distance are equal). To do this, we choose any number M greater than

the remetrized length of every path of internal edges from the root to any other internal vertex, and to a pendant edge $e = (u, v)$ we assign length

$$\tilde{\lambda}^r(e) = M - \sum_{e \in P_{r,u}} \tilde{\lambda}^r(e) > 0$$

The precise value of M will not matter in what follows so we assume some choice has been made and fixed. We refer to this remetrization as the *weighted rooted triple metrization*, due to Theorem 3.1 below.

To further elucidate the need for a choice of M, for $x, y \in X$ let

$$f_{\tilde{\lambda}^r}(x, y) = \sum_{e \in P_{r,\mathrm{MRCA}(x,y)}} \tilde{\lambda}^r(e).$$

Then $-f_{\tilde{\lambda}^r}(x, y)$ is the Gromov product (essentially the Farris transform) [DHM07] associated to $d_{\tilde{\lambda}^r}(x, y) = 2\left(M - f_{\tilde{\lambda}^r}(x, y)\right)$. For $x \neq y$, the Gromov product is independent of the choice of M, but carries all information on the topology of the tree and its internal edge lengths. However, for tree building it is convenient to pass to a tree metric, which requires a choice of M. Nonetheless, the Gromov product and the tree metric are essentially interchangable notions.

We now show the intertaxon distance $d_{\tilde{\lambda}^r}$ associated to the weighted rooted triple metrization can also be expressed in terms of information on rooted triple trees induced from T^r. For a fixed tree (T^r, λ^r) on X displaying a rooted triple $xy|z$, let $w(xy|z) = w_{\lambda^r}(xy|z)$ denote the length of the internal edge on the induced metric tree on x, y, z, which we call the *weight* of $xy|z$.

Theorem 3.1. *Suppose a rooted metric phylogenetic tree (T^r, λ^r) is given the rooted triple remetrization, $\left(T^r, \tilde{\lambda}^r\right)$. Then for all $x, y \in X$, $x \neq y$,*

$$d_{\tilde{\lambda}^r}(x, y) = 2\left(M - \sum_{xy|z \text{ on } T^r} w_{\lambda^r}(xy|z)\right),$$

where the sum is over all $z \in X$ such that $xy|z$ is displayed on T^r.

Proof. With $v = \mathrm{MRCA}(x, y)$ let $r = v_0, v_1, v_2, \ldots, v_n = v$ be the ordered nodes on the path on T^r from the root r to v, as shown in Figure 1. Let

$$k_i = |\mathrm{desc}(v_{i-1})| - |\mathrm{desc}(v_i)|,$$

the drop in number of decsendents from v_{i-1} to v_i.

For edge $e_i = (v_{i-1}, v_i)$, let $\lambda_i = \lambda(e_i)$. Then

$$\sum_{xy|z \text{ on } T^r} w_{\lambda^r}(xy|z) = \lambda_n k_n + (\lambda_n + \lambda_{n-1})k_{n-1} + \cdots + (\lambda_n + \lambda_{n-1} + \cdots + \lambda_1)k_1. \tag{2}$$

For instance the term $\lambda_n k_n$ on the right side arises because, as can be seen in Figure 1, there are k_n rooted triple trees $xy|z$, one for each z on the subtree K_n, whose internal edge length is λ_n. While Figure 1 depicts no polytomies at the v_i, the formula is valid even if there are.

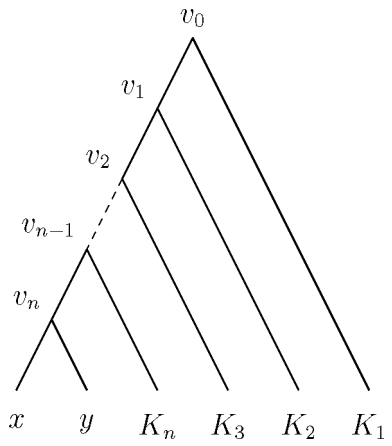

FIGURE 1. An N-taxon binary tree with root v_0 and $v_n = \mathrm{MRCA}(x, y)$. The K_i are subtrees, on k_i taxa.

Rearranging equation (2) gives

$$\begin{aligned}\sum_{xy|z \text{ on } T^r} w(xy|z) &= \lambda_n(k_n + k_{n-1} + \cdots + k_1) + \lambda_{n-1}(k_{n-1} + \cdots + k_1) + \cdots + \lambda_1 k_1 \\ &= \lambda_n \cdot n(v_n) + \lambda_{n-1} \cdot n(v_{n-1}) + \cdots + \lambda_1 \cdot n(v_1) \\ &= f_{\tilde{\lambda}^r}(x, y).\end{aligned}$$

Then by definition of $d_{\tilde{\lambda}}(x, y)$, we have

$$d_{\tilde{\lambda}^r}(x, y) = 2\left(M - f_{\tilde{\lambda}^r}(x, y)\right) = 2\left(M - \sum_{xy|z \text{ on } T^r} w(xy|z)\right),$$

as claimed. □

Example 3.2. Consider a binary rooted caterpillar tree (T^r, λ^r) on N taxa

$$(\ldots(((a_1, a_2) : \lambda_{N-2}, a_3) : \lambda_{N-3}, a_4), \ldots, a_{N-1}) : \lambda_1, a_N)$$

with the internal edges of weight $\lambda_1, \lambda_2, \ldots, \lambda_{N-2}$ from the root toward the cherry. Under the rooted triple metrization, for each a_i, a_j, $1 \le i < j \le N$,

$$\begin{aligned}f_{\tilde{\lambda}^r}(a_i, a_j) &= \sum_{e=(v,w) \in P_{r,\mathrm{MRCA}(a_i,a_j)}} \lambda^r(e) \cdot n(w) \\ &= \lambda_1 + 2\lambda_2 + \cdots + (N - j)\lambda_{N-j}.\end{aligned}$$

Also,

$$\sum_{a_ia_j|b \text{ on } T^r} w(a_ia_j|b) = \lambda_{N-j} + \left(\lambda_{N-j} + \lambda_{N-(j+1)}\right) + \cdots + \left(\lambda_{N-j} + \lambda_{N-(j+1)} + \cdots + \lambda_1\right)$$
$$= \lambda_1 + 2\lambda_2 + \cdots + (N-j)\lambda_{N-j},$$

where the terms arise from considering, in order, $b = a_{j+1}, a_{j+2}, \ldots, a_1$. Thus $f_{\tilde{\lambda}^r}(a_i, a_j) = \sum_{a_ia_j|b \text{ on } T^r} w(a_ia_j|b)$, as Theorem 2.1 showed more generally.

Example 3.3. Let $N = 2^m$ and T^r be a binary rooted balanced tree

$$(\ldots((a_1, a_2), (a_3, a_4)), \ldots, ((a_{N-3}, a_{N-2}), (a_{N-1}, a_N))\ldots)$$

on N taxa. Suppose T^r is given an equidistant metric λ^r where as one moves from the root toward any leaf the internal edge weights are in order $\lambda_1, \lambda_2, \ldots, \lambda_{m-1}$. Then edge lengths for $\left(T^r, \tilde{\lambda}^r\right)$ are

$$\tilde{\lambda}_1 = \lambda_1\frac{N}{2},\ \tilde{\lambda}_2 = \lambda_2\frac{3N}{4},\ \tilde{\lambda}_3 = \lambda_3\frac{7N}{8},\ \ldots.$$

Also, if the MRCA(a_i, a_j) is the child vertex of an edge of length λ_k, then

$$f_{\tilde{\lambda}^r}(a_i, a_j) = \sum_{e=(v,w)\in P_{r,\text{MRCA}(a_i,a_j)}} \lambda^r(e)\cdot n(w)$$
$$= \frac{N}{2}\lambda_1 + \frac{3N}{4}\lambda_2 + \cdots + N\left(1 - \frac{1}{2^k}\right)\lambda_k.$$

But also

$$\sum_{a_ia_j|b \text{ on } T^r} w(a_ia_j|b) = \underbrace{(\lambda_k + \cdots + \lambda_1) + \cdots + (\lambda_k + \cdots + \lambda_1)}_{\frac{N}{2} \text{ times}}$$
$$+ \underbrace{(\lambda_k + \cdots + \lambda_2) + \cdots + (\lambda_k + \cdots + \lambda_2)}_{\frac{N}{4} \text{ times}}$$
$$+ \cdots + \underbrace{\lambda_k + \cdots + \lambda_k}_{\frac{N}{2^k} \text{ times}}$$
$$= \frac{N}{2}\lambda_1 + \left(\frac{N}{2} + \frac{N}{4}\right)\lambda_2 + \cdots + \left(\frac{N}{2} + \frac{N}{4} + \cdots + \frac{N}{2^k}\right)\lambda_k.$$

Then $f_{\tilde{l}}(a_i, a_j) = \sum_{a_ia_j|b \text{ on } T^r} w(a_ia_j|b)$ as Theorem 2.1 demonstrated more generally.

4. Weighted Quartet metrization of an unrooted tree

For an unrooted metric tree, we define a remetrization similar to that of the last section, using weighted quarets.

Let (T, λ) be an unrooted metric tree on taxa X with $\lambda(e)$ the length of edge e. Each edge e of T determines a split (bipartition) of X, $X = M_e \sqcup N_e$, according to the taxa

on the connected components of the graph resulting from deleting e. We remetrize T by assigning to each internal edge e length

$$\tilde{\lambda}(e) = (|M_e| - 1)\,(|N_e| - 1)\,\lambda(e),$$

and to pendant edges e length $\tilde{\lambda}(e) = 1$. This gives a new metric tree $\left(T, \tilde{\lambda}\right)$, which we refer to as having the *weighted quartet metrization*, due to Theorem 4.2 below. The distance between x and y on the remetrized tree is

$$d_{\tilde{\lambda}}(x, y) = 2 + \sum_{e \in P_{x,y}} (|M_e| - 1)\,(|N_e| - 1)\,\lambda(e).$$

We will show this intertaxon distance can also be expressed in terms of information from quartet trees induced from T. As a first step, for a quartet Q let $E(Q)$ denote the set of edges on the path in T which induces the internal edge of the quartet tree, and $N(e; x, y)$ be the number of quartets $Q \in \mathcal{Q}_{x,y}$ for which $e \in E(Q)$. Then

$$N(e; x, y) = \sum_{\substack{Q \in \mathcal{Q}_{x,y} \\ e \in E(Q)}} 1. \tag{3}$$

Lemma 4.1. *Let T be an unrooted metric phylogenetic tree on taxa X. Then for all $x, y \in X$, $x \neq y$, and internal edges $e \in P_{x,y}$*

$$N(e; x, y) = (|M_e| - 1)\,(|N_e| - 1) \tag{4}$$

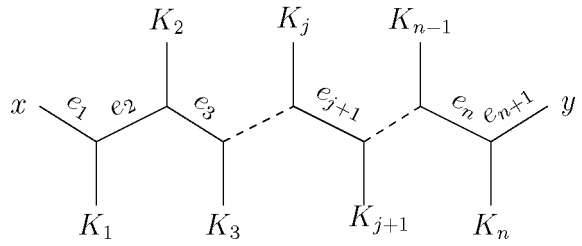

FIGURE 2. The path between taxa x and y on an N-taxon unrooted binary metric tree. The K_i represent subtrees.

Proof. Let $P_{x,y} = \{e_1, \ldots, e_{n+1}\}$ and $M_i|N_i$ be the split on T associated to e_i. If the path from x to y contains no polytomies, from Figure 2 we see by equation (3) that if k_i denotes the number of taxa on the subtree K_i then

$$\begin{aligned} N(e_2; x, y) &= k_1k_2 + k_1k_3 + \cdots + k_1k_n = k_1\,(k_2 + k_3 + \cdots k_n) \\ N(e_3; x, y) &= (k_1k_3 + \cdots + k_1k_n) + (k_2k_3 + + \cdots + k_2k_n) \\ &= (k_1 + k_2)\,(k_3 + \cdots + k_n)\,, \end{aligned}$$

and more generally

$$N(e_i; x, y) = \left(\sum_{j=1}^{i} k_j\right)\left(\sum_{j=i+1}^{k+1} k_j\right) = (|M_{e_i}| - 1)\,(|N_{e_i}| - 1)\,,$$

as claimed. If there are polytomies along the path from x to y, one readily sees the same formula applies. □

For a fixed tree (T, λ) on X displaying a quartet $Q = xy|zv$, let $w(Q) = w_\lambda(Q)$ denote the length of the internal edge on the induced metric tree on x, y, z, v, which we call the *weight* of Q.

Theorem 4.2. *Let (T, λ) be an unrooted, binary metric tree on X with $x, y \in X$. Then*

$$d_{\tilde{\lambda}}(x, y) = 2 + \sum_{Q \in \mathcal{Q}_{x,y}} w(Q).$$

Proof. By definition of $w(Q)$, we have $w(Q) = \sum_{e \in E(Q)} \lambda(e)$. Then

$$\begin{aligned}
2 + \sum_{Q \in \mathcal{Q}_{x,y}} w(Q) &= 2 + \sum_{Q \in \mathcal{Q}_{x,y}} \sum_{e \in E(Q)} \lambda(e) \\
&= 2 + \sum_{e \in P_{x,y}} \lambda(e) \sum_{\substack{Q \in \mathcal{Q}_{x,y} \\ e \in E(Q)}} 1 \\
&= 2 + \sum_{e \in P_{x,y}} \lambda(e) N(e; x, y) && \text{by equation (3),} \\
&= 2 + \sum_{e \in P_{x,y}} \lambda(e)\,(|M_e| - 1)\,(|N_e| - 1) && \text{by Lemma 4.1,} \\
&= d_{\tilde{\lambda}}(x, y).
\end{aligned}$$

□

Example 4.3. The unrooted 8−taxon caterpillar tree

$$(T, l) = (\ldots(((a_1, a_2) : \lambda_1, a_3) : \lambda_2, a_4), \ldots, a_6) : \lambda_5, a_7), a_8),$$

shown in Figure 3, when remetrized with the quartet metrization $\tilde{\lambda}$ has internal edges of weight

$$(1 \cdot 5)\,\lambda_1, (2 \cdot 4)\,\lambda_2, (3 \cdot 3)\,\lambda_3, (4 \cdot 2)\,\lambda_4, (5 \cdot 1)\,\lambda_5,$$

and pendant edges of length 1.

Let $x = a_3$ and $y = a_6$. Then we have

$$d_{\tilde{\lambda}}(a_3, a_6) = 2 + (2 \cdot 4)\,\lambda_2 + (3 \cdot 3)\,\lambda_3 + (4 \cdot 2)\,\lambda_4.$$

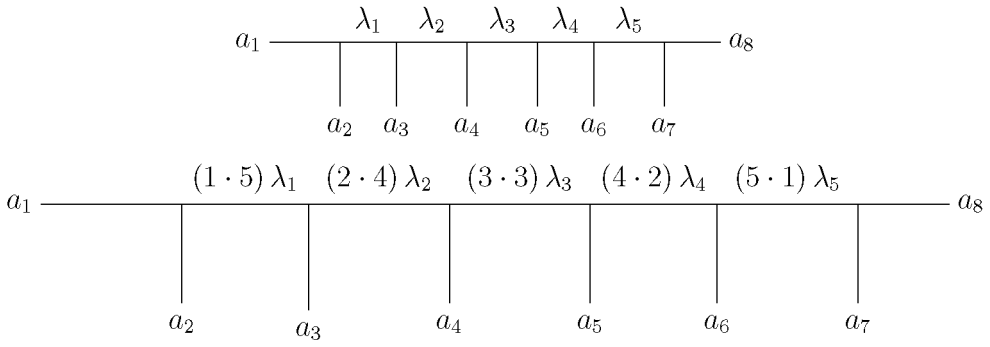

FIGURE 3. An 8−taxon metric caterpillar tree (T, λ) (top) and its quartet remetrization $\left(T, \tilde{\lambda}\right)$ (bottom).

The 13 quartet trees on T separating a_3 and a_6 are shown in Figure 4, so

$$\begin{aligned}\sum_{Q \in \mathcal{Q}_{a_3,a_6}} w(Q) &= 2 \cdot \lambda_2 + 1 \cdot \lambda_3 + 2 \cdot \lambda_4 + 2 \cdot (\lambda_2 + \lambda_3) + 2 \cdot (\lambda_3 + \lambda_4) + 4 \cdot (\lambda_2 + \lambda_3 + \lambda_4) \\ &= (2 \cdot 4)\,\lambda_2 + (3 \cdot 3)\,\lambda_3 + (4 \cdot 2)\,\lambda_4 \\ &= d_{\tilde{\lambda}}(a_3, a_6),\end{aligned}$$

as Theorem 4.2 states.

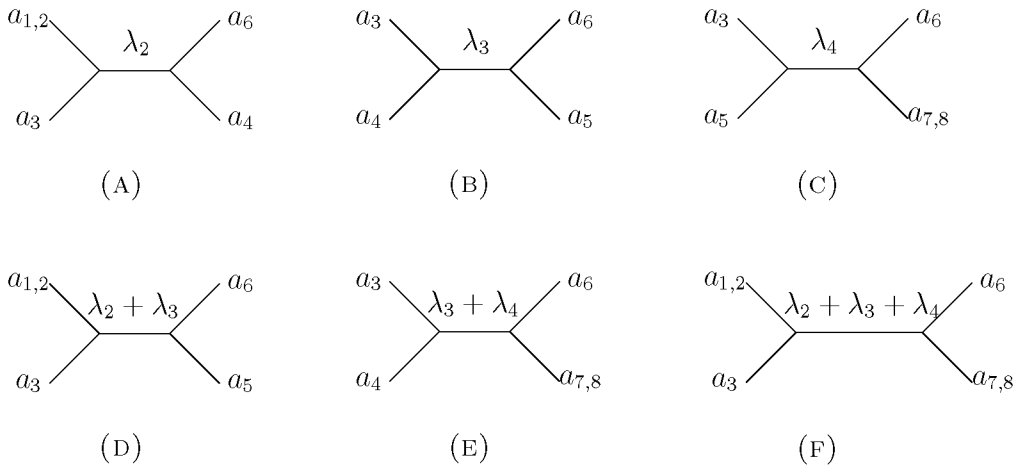

FIGURE 4. The 13 quartet trees on (T, λ) separating a_3 and a_6. Multiple taxa on a leaf represent choices leading to multiple quartet trees.

Example 4.4. Consider an unrooted balanced tree

$$(((a_1, a_2) : \lambda_1, (a_3, a_4) : \lambda_2) : \lambda_3, ((a_5, a_6) : \lambda_4, (a_7, a_8) : \lambda_5))$$

on 8 taxa as shown in Figure 5. After remetrization, we have internal edges of weight

$$(1 \cdot 5)\,\lambda_1, (1 \cdot 5)\,\lambda_2, (3 \cdot 3)\,\lambda_3, (1 \cdot 5)\,\lambda_4, (1 \cdot 5)\,\lambda_5.$$

Suppose $x = a_3$ and $y = a_6$. Then

$$d_{\tilde{\lambda}}(a_3, a_6) = (1 \cdot 5)\,\lambda_2 + (3 \cdot 3)\,\lambda_3 + (1 \cdot 5)\,\lambda_4.$$

On the other hand, by listing the 13 quartet trees separating a_3 and a_6 we find:

$$\begin{aligned}\sum_{Q \in \mathcal{Q}_{a_3, a_6}} w(Q) &= 2 \cdot \lambda_2 + 4 \cdot \lambda_3 + 2 \cdot \lambda_4 + 2 \cdot (\lambda_2 + \lambda_3) + 2 \cdot (\lambda_3 + \lambda_4) + 1 \cdot (\lambda_2 + \lambda_3 + \lambda_4) \\ &= (2 + 2 + 1)\lambda_2 + (4 + 2 + 2 + 1)\lambda_3 + (1 + 2 + 1)\lambda_4,\end{aligned}$$

which is equal to $d_{\tilde{\lambda}}(a_3, a_6)$.

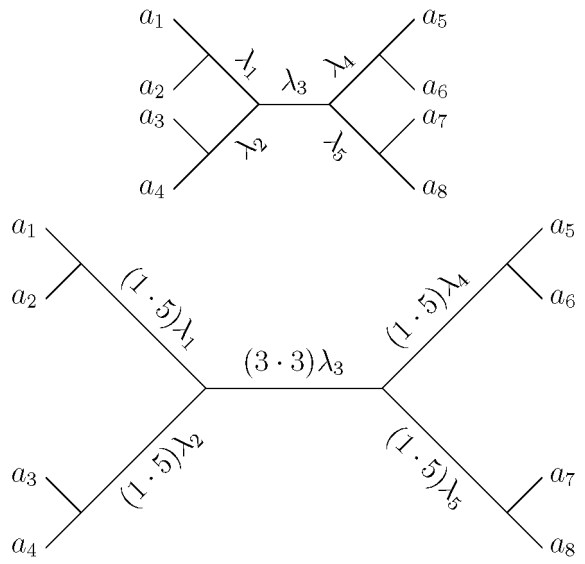

FIGURE 5. An unrooted 8-taxon balanced metric tree, with original edge lengths (top) and quartet remetrization (bottom).

5. Weighted Quartet Distance Supertree and Consensus Algorithms

Since, by Theorems 3.1 and 4.2, the pairwise distances between taxa on trees given the rooted triple or quartet remetrizations of the previous sections can be computed from knowing only the weighted rooted triples or weighted quartets displayed on the original tree, they lead to new methods of inferring a large metric tree from that information.

After computing pairwise distances from weighted rooted triples or quartets using the formulas of Theorem 3.1 or 4.2, a standard distance-based tree construction algorithm can be used to build the remetrized tree. Then the individual internal edge lengths can be adjusted to remove the multiplier arising from the tree topology in the remetrization. If the tree construction method is robust to some noise, then the presence of a sufficiently small number of erroneous quartets, or sufficiently small errors in the weights, should still allow for construction of an approximation to the original metric tree, with pendant edges weights set to 1.

5.1. **Inferring a tree from displayed weighted quartets.** In the quartet case, we present this as a formal algorithm. Let $\mathcal{M}$ denote any method of constructing a metric tree from pairwise distances between taxa. For example for $\mathcal{M}$ one might choose Neighbor Joining (NJ) [SK88] or FastME [LDG15].

Algorithm 5.1. (WQDS/$\mathcal{M}$) Weighted Quartet Distance Supertree with method $\mathcal{M}$
Input: A collection $\mathcal{Q}$ of weighted quartets on taxa in X

(1) For each pair $x, y \in X$ of taxa, $x \neq y$, with $\mathcal{Q}_{x,y} \subset \mathcal{Q}$ the subset of weighted quartets separating x and y, define the distance

$$d_{\tilde{\lambda}}(x, y) = 2 + \sum_{Q \in \mathcal{Q}_{x,y}} w(Q).$$

(2) Use the distance method $\mathcal{M}$ to build an unrooted metric tree $(T, \tilde{\lambda})$ from $d_{\tilde{\lambda}}$.
(3) For each internal edge e on T with associated split $M_e|N_e$, let

$$\lambda(e) = \frac{\tilde{\lambda}(e)}{(|M_e| - 1)(|N_e| - 1)}.$$

For pendant edges e, let $\lambda(e) = 1$.

Output: An unrooted metric tree (T, λ) on X.

The first step of this algorithm, when applied to a set composed of one weighted quartet per choice of 4 taxa in X has running time $\mathcal{O}(|X|^4)$: One must consider $\binom{|X|}{4}$ quartets, each of which contributes to 4 of the $\binom{|X|}{2}$sums in that step. If $\mathcal{M}$ is NJ, the second step requires time $\mathcal{O}(|X|^3)$ to obtain a metric tree. By traversing the edges of the tree once, one can compute the M_e, N_e and adjust the edge lengths as in step 3, for an additional time of $\mathcal{O}(|X|)$. Thus the entire algorithm is accomplished in time $\mathcal{O}(|X|^4)$.

For WQDS to be used, its input of weighted quartet trees must first be obtained. For one genetic locus one might, for example, infer all metric quartet trees on X by standard phylogenetic methods, and use the resulting weighted quartets. However, as direct inference of large trees for one locus is already well established and relatively quick, and older quartet methods for this problem are no longer in use, we do not further explore that application. Instead we consider a problem of greater current interest: inferring a species tree from a collection of gene trees.

5.2. **Inferring a species tree from gene trees.** The standard model for the generation of gene trees from a fixed metric species tree is the *multispecies coalescent model* (MSC) [PN88]. The species tree, denoted by σ^r, is rooted with edge weights in *coalescent units*. Coalescent units are obtained from more biologically natural units by inversely scaling the number of generations the edge represents, by the population size, as these cannot be separately identified under the MSC. If the population size is a constant N and the edge represents t generations, the edge weight is simply t/N. If the population varies with time $s \in [0, t]$ along the edge, then the weight is

$$\int_0^t \frac{1}{N(s)}\, ds.$$

Under the MSC with one sampled gene lineage per taxon, if the species tree σ^r displays a quartet $ab|cd$ with weight x (the length of the induced quartet tree's internal edge in coalescent units), then the probabilities that a gene tree will display each of the three resolved quartet topologies on these taxa are [ADR11]

$$p_{ab|cd} = 1 - \frac{2}{3}\exp(-x),\ p_{ac|bd} = \frac{1}{3}\exp(-x),\ p_{ad|bc} = \frac{1}{3}\exp(-x).$$

If the rooted triple $ab|c$ with weight x is displayed on σ^r, then the same formulas give probabilities of a gene tree displaying rooted triples $ab|c$, $ac|b$, and $bc|a$ respectively [PN88]. In particular, since $x > 0$, the quartet or rooted triple with the highest probability of being displayed on a gene tree is the one displayed on the species tree.

This suggests the following algorithm for inferring an unrooted metric species tree from a collection of gene trees under the MSC.

Algorithm 5.2. (WQDC/$\mathcal{M}$) Weighted Quartet Distance Consensus with method $\mathcal{M}$
Input: A collection of n topological gene trees on taxa X

(1) For each subset of four taxa $x, y, z, w \in X$, determine the counts of the quartets $xy|zw$, $xz|yw$, and $xw|yz$ displayed on the gene trees.
(2) For each subset of four taxa $x, y, z, w \in X$, choose the dominant (i.e, most frequent) quartet as the estimated quartet topology. In the case of a tie, choose from the most frequent uniformly at random. With n_{dom} the number of gene trees displaying the dominant quartet on x, y, z, w, solve the equation

$$1 - \frac{2}{3}e^{-\hat{x}} = \frac{n_{dom}}{n}$$

to find $\hat{x}$ as the estimated weight of the dominant quartet tree.
(3) Apply WQDS/$\mathcal{M}$ to the set of $\binom{n}{4}$ estimated weighted dominant quartets.

Output: An unrooted metric tree on X

As discussed in [Rho19], step (1)(a) can be accomplished in time $\mathcal{O}(|X|^4 n)$, with step (2) requiring only time $\mathcal{O}(|X|^4)$. Combined with the time for WQDS/$\mathcal{M}$ for $\mathcal{M}$=NJ shown earlier, the total time is $\mathcal{O}(|X|^4 n)$. Thus, the most time intensive step in the algorithm is tallying the displayed quartets.

Let us say a distance method $\mathcal{M}$ of constructing a metric tree from pairwise distances is *well-behaved* if 1) when applied to a tree metric returns the unique tree it fits, and 2) is continuous at all tree metrics. The second requirement means that a sufficiently small perturbation in a distance table fitting a binary tree will result in an output of the same binary tree topology, with only small perturbations in the edge weights. Both NJ and Minimum Evolution (ME) are well-behaved, though in practice the heuristic FastME is often used in place of ME.

Theorem 5.3. *Let $\mathcal{M}$ be any well-behaved distance method for tree building. Under the MSC model with one sampled lineage per taxon per gene, on a binary rooted metric species tree (σ^r, λ^r), the output of the WQDC/$\mathcal{M}$ algorithm is a statistically consistent estimator of both the unrooted topological tree σ and the internal edge lengths in λ.*

Proof. Consider a collection of n gene trees generated under the MSC on (σ^r, λ^r). Then for each choice of four taxa x, y, z, w, by the law of large numbers as $n \to \infty$ the probability that the dominant quartet topology matches the quartet displayed on the species tree $\to 1$. Similarly, for any choice of $\epsilon > 0$ the probability that the estimated weight $\hat{x}$ is within ϵ of the quartet weight on the species tree also $\to 1$. Since there are a finite number of sets of 4 taxa, as $n \to \infty$ the probability that all dominant quartet topologies match that on the species tree, and all weights are within ϵ of the true value also $\to 1$.

Thus for any choice of $\epsilon > 0$, with probability $\to 1$ as $n \to \infty$ the computed pairwise quartet distances will be within ϵ of the true values on the species tree with the quartet remetrization. Since $\mathcal{M}$ is well behaved, with probability $\to 1$ it will return the unrooted topology of σ, with internal edge lengths differing from true remetrized values by arbitrarily small amounts. Adjusting the lengths of the internal edges to estimate the original species tree edge lengths involves dividing by a number ≥ 1, so as $n \to \infty$ these estimates can also be made within ϵ of the true values with probability 1. □

It is actually not necessary that all taxa in X are on all gene trees for statistical consistency. As was done in [Rho19] for the method QDC, one can relax that condition as long as 1) the pattern of missingness of taxa is independent of the gene tree topology, and 2) as the total number of gene trees goes to infinity, so does the number on which each set of 4 taxa appears.

Note that WQDC/$\mathcal{M}$ as presented above does not allow for inference of pendant edge weights on σ. However, if input gene trees have at least 2 samples per taxon, one can infer those as well, by simply considering an extended species tree obtained by appending two edges of length 0 to each leaf. Similar modifications allow for more samples per taxon.

Remark 5.4. For Weighted Rooted Triple Distance Supertree with method $\mathcal{M}$ (WRTDS/$\mathcal{M}$), one replaces the formulas in steps (1) and (3) of Algorithm 5.2 with similar one arising from Theorem 3.1 and equation (1). Note that $\mathcal{M}$ can now be chosen to assume ultrametricity of the distance (e.g., UPGMA), since $d_{\tilde{l}}$ approximates an ultrametric tree metric. If such an $\mathcal{M}$ is used, then a rooted tree will be returned, and an estimate of both the rooted topology and all its internal edges will be inferred.

Weighted Rooted Triple Distance Consensus with method $\mathcal{M}$ (WRTDC/$\mathcal{M}$) is given by modifying Algorithm 5.2 to count displayed rooted triples, and use WRTDS/$\mathcal{M}$.

A consistency result for WRTDC/$\mathcal{M}$ can be shown similarly to Theorem 5.3.

Remark 5.5. In applying WQDC/$\mathcal{M}$ to data, there is one serious practical issue that may need to be addressed. In a finite sample of gene trees, one may find that the dominant quartet for a set of 4 taxa is displayed on every gene tree. Then solving

$$1 - \frac{2}{3}e^{-x} = 1$$

leads to an estimated weight of ∞ for that quartet. While this correctly indicates the weight should be large, it does not give the finite estimate that is typically needed for applying a tree building method.

Since the MSC does not give expected counts of 100% for one quartet topology for any finite edge weight, this situation can be interpreted as a sign of an insufficient number of gene trees in the data set to properly estimate the weight. One approach to addressing this is to treat counts of $(n, 0, 0)$ for the 3 topologies on a given set of 4 taxa as having dominant count $n - 1/2$ out of a total of n. That is, we reduce the actual count slightly, by less than 1, to represent an expected count that our sample size would still be likely to show as 100% agreement.

This *ad hoc* adjustment will result in all infinite weights being replaced by the same finite number. But note that with such weights need not result in a good approximation to the desired distance between taxa. A better approach, though one that may not be feasible given practical data collection constraints, is simply to obtain more gene trees so this situation does not occur, or restrict to collections of taxa that are closely enough related so that all sets of 4 taxa show some quartet discordance across the gene trees.

As will be shown through simulations in the next section, WQDC/$\mathcal{M}$ may not perform as well as other methods for inferring the topology of the species tree. The reason for this appears to be our inability to obtain accurate estimates of the weight of quartets when they are displayed on all, or almost all, gene trees. While the heuristic described above gives us a finite estimate which is necessary to have the finite distances between taxa that the algorithm requires, it is unlikely to be very accurate. Even if a handful of gene trees display a quartet other than the dominant one, the estimate of the weight is often not to very accurate.

This is not an unusual situation as it often occurs when four taxa are widely placed on a species tree, and can occur for taxa whose displayed quartet has only a single edge of the species tree as its internal edge, provide that edge is long in coalescent units. However, simulations suggested to us that a tree inferred by WQDC/$\mathcal{M}$ often did correctly display many correct splits, and those with long edge lengths tended to be correct. That observation is the basis for the following algorithm. It proceeds by using WQDC/$\mathcal{M}$ to pick only one split on the species tree with the largest weight, then dividing the taxa into two groups by this split, and recursively building subtrees on these groups. This process seeks to divide the taxa into smaller groups that will be closer together, so that the poor behavior caused by long edges will not be present in the later stages of the recursion. While it cannot be expected to improve edge length estimates of longer edges, the hope is that the shorter lengths will be estimated well.

Algorithm 5.6. (Recursive WQDC/$\mathcal{M}$) Recursive Weighted Quartet Distance Consensus with method $\mathcal{M}$

Input: A collection of n topological gene trees on taxa X, and positive number L

(1) For each subset of four taxa $x, y, z, w \in X$, Determine the counts of the quartets $xy|zw$, $xz|yw$, and $xw|yz$ displayed on the gene trees.

(2) If X has 3 or fewer taxa, return the unique unrooted tree on X with all edge lengths 1. Otherwise,

  (a) Apply Steps (2) and (3) of WQDC/$\mathcal{M}$ to the quartet counts obtain an estimated metric species tree τ.

  (b) If all internal edge weights on τ are less than L, return τ.

  (c) Let $X_0|X_1$ be the split of X associated to the longest edge of τ, and $\ell_{X_0|X_1}$ its length. In the case of a tie, choose the edge uniformly at random from the longest edges.

  (d) Create taxon sets $X_0' = X_0 \cup \{y_1\}$ and $X_1' = X_1 \cup \{y_0\}$, where y_0, y_1 represent "composite taxa" for the split sets X_0, X_1. For each choice of 4 taxa in X_i' compute quartet counts as follows: For quartets containing y_{1-i}, sum over $x \in X_{1-i}$ the counts from Step (1) containing x in place of y_{1-i}. For quartets containing only elements in X_i, retain the quartet counts from Step (1).

  (e) Recursively apply Step (2) to the quartet counts for X_0' and X_1' to obtain metric trees τ_0, τ_1 on X_0', X_1'.

  (f) Form a metric tree σ by identifying leaf y_1 on τ_0 with y_0 on τ_1, surpressing that node, and assigning the conjoined edge length $\ell_{X_0|X_1}$. Return σ.

Output: An unrooted metric tree on X

Step (1) requires time $\mathcal{O}(|X|^4 n)$. One application of Step (2) (without the recursive call) on quartet counts for k taxa has time $\mathcal{O}(k^4)$. In the worse case, the split sets have sizes 2,$k-2$ for each recursive call and at every step there is an internal edge weights $\geq L$, leading to time $\mathcal{O}(|X|^4 n + |X|^5)$ for the entire algorithm. However, variations on this algorithm, in which all splits with weights over L in the tree of Step (1) are retained might reduce the typical running time considerably in practical use.

A reasonable choice for the parameter L might be $L = 2$. This corresponds to the quartets defining an edge of length < 2 having an expected frequency of at most $1 - (2/3)\exp(-2) \approx 0.9098$ of the displayed gene quartets matching the species tree quartet.

6. Algorithm Performance in Simulations

Although the algorithms of the last section provide statistically consistent estimators of a species tree from gene trees under the MSC model, their practical performance will be affected by several factors. First, even if gene trees are sampled from the MSC with no error, an algorithm cannot be expected to always infer the underlying species tree from a finite sample of gene trees. Second, if the input gene trees for the algorithm are inferred from sequences that were simulated along the gene trees under some standard substitution model, there is likely to be some inference error in the gene trees due to the finiteness of sequences. Finally, for empirical data neither the MSC nor the substitution

models may exactly describe the true processes, so that there is additional error from model misspecification. Although the performance of phylogenetic inference methods under model misspecification is rarely investigated, simulations can provide insight into the effects of the first two issues.

As an initial, and limited, investigation into the performance of the algorithms of the last section, we present some simulation analysis following the framework of [Rho19], using the simulated Avian data sets of [BMW14] which were also used in [VW15]. All calculations were performed in R using the `ape` [PS18] and `MSCquartets` [ABMR19] packages. These data sets for a fixed species tree contain both a sample of gene trees under the MSC, and inferred gene trees from sequences simulated on the sampled gene trees. In addition, there are similar datasets for rescalings of the species tree by factors of 0.5 and 2, to respectively increase and decrease the amount of incomplete lineage sorting. For details on the simulation and gene tree inference procedure, see the referenced publications.

To reduce computation time, we pass from the original 48-taxon species tree, to the 30-taxon subtree described in [Rho19]. We similarly pass to subtrees of both gene trees sampled under the MSC, and subtrees of inferred gene trees. Although these subtrees of inferred gene trees may not be exactly the trees that would be inferred from the subset of sequences, differences are likely to be small.

We quantify the accuracy of methods in two ways. First, for topological accuracy, we compute the normalized Robinson-Foulds (RF) distance between the true unrooted species tree and the inferred one. The normalization is such that two trees displaying none of the same non-trivial splits will have distance 1, and two binary 30-taxon trees differing by a single NNI move have distance $2/2(30-3) = 0.037$.

Second, for metric accuracy, we use a non-standard variant of a distance of Kuhner and Felsenstein [KF94] between the true species tree and the inferred one. For the KF distance as implemented in `ape`, for each tree one first forms a vector whose entries correspond to all possible splits of the taxa, with an entry of the length of the edge defined by the split if it is displayed on the tree and 0 otherwise. The Euclidean distance between the vectors for the two trees then gives the distance. The variant we use, denoted $KF[x]$, replaces any vector entry corresponding to a trivial split with 1 and any entry larger than x with x. This treatment of trivial splits is necessary since pendant edge lengths cannot be inferred from this data. The treatment of entries larger than x prevents a split that is displayed on both trees with defining edges of length $\geq x$ but of significantly different size from influencing the distance. We use $x = 2$ here, since such long edges on the true species tree give rise to expected quartet counts in which one is large and the others small. These are precisely the counts for which stochastic variation produces large variation in estimated lengths. Note that if two trees differed by splits with edge lengths ≥ 2, then their $KF[2]$ distance would be at least 4. Thus a $KF[2]$ distance less than 4 indicates the two tree topologies agree on all long-edge splits. The choice of 2 here is of course arbitrary, but based on the reasoning given at the end of the last section.

The simulated data sets contain 20 replicates of 1000 sampled and inferred gene trees for each condition, with gene trees inferred from 500 base sequences. In Figures 6 and 7 we illustrate the mean over the replicates of the distances of inferred species trees from

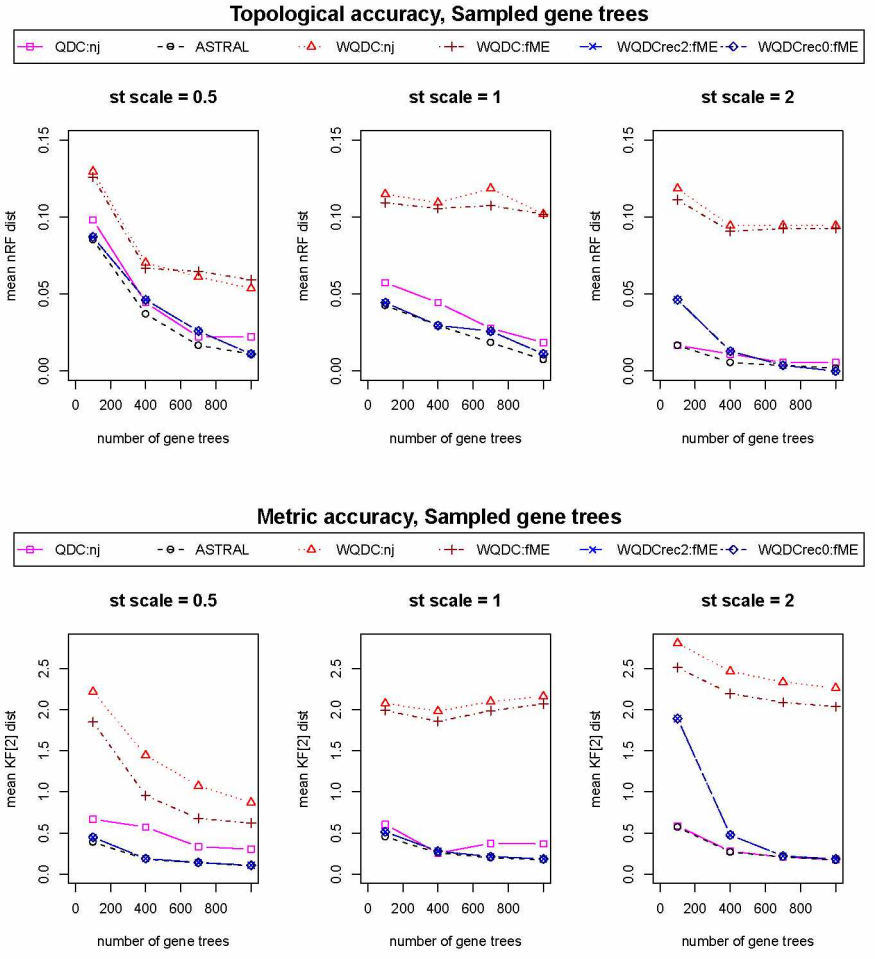

FIGURE 6. Simulation results on accuracy of methods of inference of species trees from gene trees sampled under the MSC.

the true one. Results are given for $g = 100$, 400, 700, and 1000 gene trees, by using only the first g gene trees in each simulated collection. We present results of WQDC (Algorithm 5.2) using both the NJ and fastME algorithms for tree building, as well as Recursive WQDC (Algorithm 5.6) using FastME for $L = 2, 0$. For comparison to other methods, we include ASTRAL and QDC, which were already compared for topological inference in [Rho19]. Internal edge lengths for trees inferred by these methods, which infer only topological trees, were assigned by methods that use only counts of quartets for sets of four taxa defining those edges, see [ZRSM18, ABMR19] for precise desciptions.

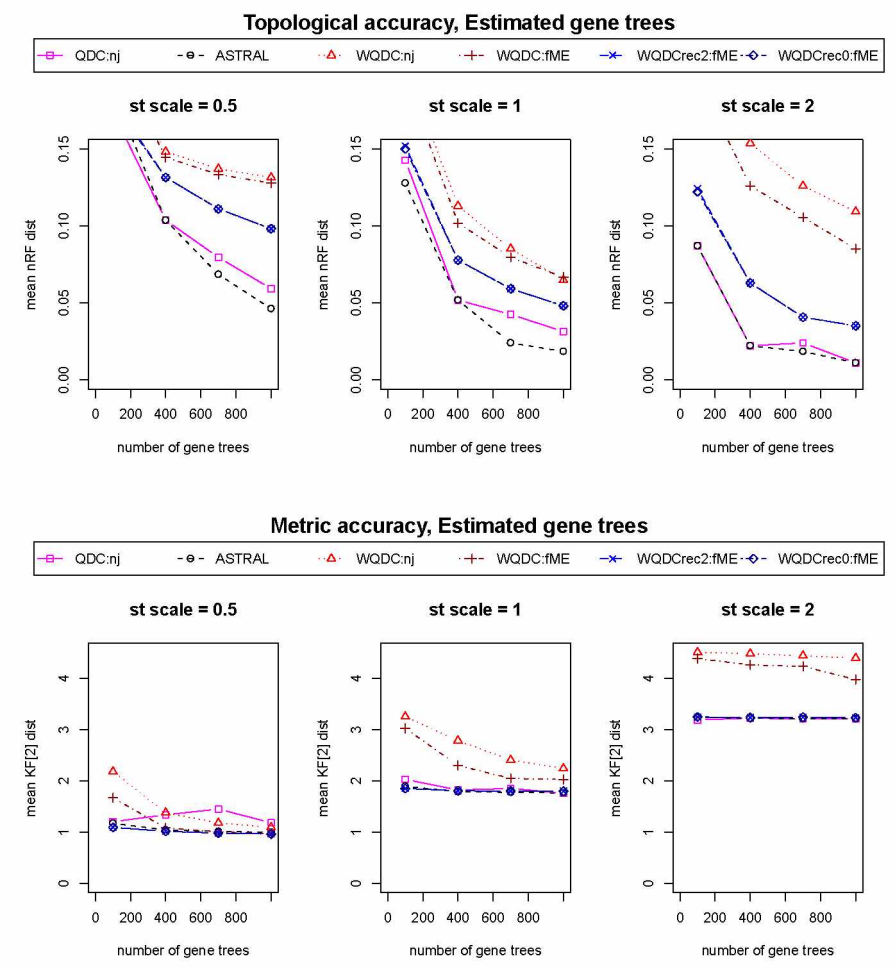

FIGURE 7. Simulation results on accuracy of methods of inference of species trees from gene trees inferred from sequences simulated on trees sampled under the MSC.

For gene trees sampled from the MSC, with no inference error, Figure 6 indicates that WQDC, with either distance method, has considerably poorer topological and metric accuracy than the other methods used. While Figure 7 shows similar results for the methods applied to inferred gene trees, the gap in performance between these methods and others is narrowed. The recursive WQDC, with $L = 0$ or 2 offers a clear improvement over non-recursive WQDC in all situations. This suggests that the source of the poor performance of the non-recursive WQDC is indeed the poor estimation of long edge lengths, as the recursive algorithm operates in such a way that after splits for such edges are put into the tree being inferred, the length of those edges no longer influences future steps. Finally,

since there is no substantial difference in the performance of the recursive WQDC for $L = 0$ and $L = 2$, it appears only long edges degrade performance. Since larger values of L reduce running time, this can have an impact for practical use.

When compared to QDC or ASTRAL, the recursive WQDC's performance is usually worse. For topological accuracy, the normalized RF distance is, however, generally less than the 0.037 a single NNI move produces for a 30-taxon tree, so the difference is not great. Interestingly, for metric accuracy, recursive WQDC often matches the best performing algorithm.

Nonetheless, on this one set of simulations ASTRID gives the best topological and metric accuracy among all these quartet-based method. This suggests that if either variant of WQDC is to be useful for empirical inference of species trees, additional development will be needed. We note that while its unweighted analog QDC also is slightly outperformed by ASTRID, it nonetheless serves as a crucial building block to the NANUQ algorithm for network inference [ABR19], which does have several practical advantages over other network inference methods. There may be similar roles for WQDC.

Acknowledgements

This work was supported by the National Institutes of Health grant R01 GM117590, awarded under the Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological and Mathematical Sciences.

References

[ABMR19] E.S. Allman, H. Baños, J.D. Mitchell, and J.A. Rhodes. *MSCquartets: Analyzing Gene Tree Quartets under the Multi-Species Coalescent*, 2019. R package version 1.0.5.

[ABR19] E.S. Allman, H. Baños, and J.A. Rhodes. NANUQ: a method for inferring species networks from gene trees under the coalescent model. *Algorithms Mol. Biol.*, 14(24):1–25, 2019.

[ADR11] E.S. Allman, J.H. Degnan, and J.A. Rhodes. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *J. Math. Biol.*, 62(6):833–862, 2011.

[ADR13] E.S. Allman, J.H. Degnan, and J.A. Rhodes. Species tree inference by the STAR method, and generalizations. *J. Comput. Biol.*, 20(1):50–61, 2013.

[ADR18] E.S. Allman, J.H. Degnan, and J.A. Rhodes. Species tree inference from gene splits by Unrooted STAR methods. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 15:337–342, 2018.

[BMW14] M.S. Bayzid, S. Mirarab, and T. Warnow. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. *PLOS One*, 2014.

[DE03] A.W.M. Dress and P.L. Erdős. X-trees and weighted quartet systems. *Ann. Comb.*, 7(2):155–169, 2003.

[DHM07] A. Dress, K.T. Huber, and V. Moulton. Some uses of the Farris transform in mathematics and phylogenetics—a review. *Ann. Comb.*, 11(1):1–37, 2007.

[GHMS08] S. Grünewald, K.T. Huber, V. Moulton, and C. Semple. Encoding phylogenetic trees in terms of weighted quartets. *J. Math. Biol.*, 56(4):465–477, 2008.

[HD10] J. Heled and A.J. Drummond. Bayesian inference of species trees from multilocus data. *Mol. Biol. and Evol.*, 27(3):570–580, 2010.

[KF94] M.K. Kuhner and J. Felsenstein. Simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Bio. Evol.*, 11:459–468, 1994.

[LDG15] V. Lefort, R. Desper, and O. Gascuel. FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.*, 32(10):2798–2800, 2015.

[Liu08] L. Liu. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, 24(21):2542–3, 2008.

[LY11] L. Liu and L. Yu. Estimating species trees from unrooted gene trees. *Syst. Biol.*, 60:661–667, 2011.

[LYPE09] L. Liu, L. Yu, D.K. Pearl, and S.V. Edwards. Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.*, 58:468–477, 2009.

[PN88] P. Pamilo and M. Nei. Relationships between gene trees and species trees. *Mol Biol Evol.*, 5(5):568–83, 1988.

[PS18] E. Paradis and K. Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528, 2018.

[Rho19] J.A. Rhodes. Topological metrizations of trees, and new quartet methods of tree inference. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, pages 1–12, 2019. early access.

[SK88] J. Studier and K. Keppler. A note on the Neighbor-Joining algorithm of Saitou and Nei. *Mol. Bio. Evol.*, 5:729–731, 1988.

[VW15] P. Vachaspati and T. Warnow. ASTRID: Accurate Species TRees from Internode Distances. *BMC Genomics*, 16(Suppl 10):S3, 2015.

[ZRSM18] C. Zhang, M. Rabiee, E. Sayyari, and S. Mirarab. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19 (Suppl 6)(153):15–30, 2018.

Department of Mathematics and Statistics, University of Alaska Fairbanks, 99775
E-mail address: syourdkhani@alaska.edu,j.rhodes@alaska.edu

Chapter 3: Identifiability of a Protein Model[2]

3.1 Abstract

The Profile Mixture Model is used to analyze protein sequence data in which various sites are suspected to follow many different substitution processes on a single evolutionary tree. A fundamental question for such a complex and highly-parameterized model is whether the parameters are identifiable. In this work this question is answered positively. The main result shows that for a tree with more than 8 taxa, both the tree topology and all numerical parameters are generically identifiable when the number of profiles is less than 77.

3.2 Introduction

If we have protein sequences for an ancestral species and its descendants, by looking at the each site, we can see that the substitution process may vary from one site to another. Sometimes the rate of change is fast, sometimes slow, and sometimes no change occurs. This behavior suggests defining different classes of substitution processes for different sites and leads to the concept of mixture models that will be explained in Section 3.3. These models may provide a better fit to data, though they have many more parameters than more standard single-class models. But if a mixture model is used in data analysis, it is important that the model's parameters are identifiable. That means parameter values can be recovered from the expected joint distribution, and thus a researcher can hope to recover parameter values from data.

Allman and Rhodes [1] proved that for a general Markov (GM) mixture model when the number of classes is less than the number of states, the 4-leaf species tree topology is identifiable for DNA sequences. This result immediately applies to larger trees as well. In another work, Rhodes and Sullivant [5] showed that for an r-component identical tree mixture of the GM model of character evolution with κ-state random variables on an n-leaf

[2]This chapter is being prepared to be published as a joint publication with E. S. Allman and J. A. Rhodes.

binary phylogenetic tree, under mild technical conditions, both the tree parameter and the numerical parameters are generically identifiable.

The model we consider in the following chapters is called the *Profile Mixture (PM) Model*, introduced by Quang et al. [4]. The question answered in this work is whether the parameters of this model, used to infer phylogenetic trees from protein sequence data, are identifiable. In this model, there are 20 states, corresponding to 20 amino acids, and the number of classes we are interested in is more than 60. Some parameters for this model include the 60 rates of evolution of each class, and a vector giving the relative sizes of the 60 rate classes. Other parameters in this model are: a tree topology with n taxa and $2n-3$ edge lengths, entries of a 20×20 symmetric matrix giving "relative substitution rates", and for each class 20 entries of the vector of equilibrium state frequencies. Because of the complexity of the model, determining identifiability is difficult. Unfortunately, the results mentioned earlier on the GM model do not apply, since while the PM model is a submodel of the GM, it is not generic within it.

The proof strategy we follow employs algebraic concepts. We use tensors to represent the probability distribution in a phylogenetic model and then Kruskal's theorem to identify components of tensors uniquely. We can borrow from algebraic geometry the idea that we can extend an identifiability result for a special choice of parameters to generic choices of phylogenetic models.

Section 3.3 will more carefully introduce phylogenetic substitution models, and in particular the Profile Mixture Model. Section 3.4 provides the algebraic definitions and lemmas we use, though removed from the biological setting of interest. Section 3.5 connects the phylogenetic model we study with algebraic notions. Then in Section 3.6 the proof of our main theorem on identifiability of the PM model appears.

3.3 Markov Models on Trees

In this chapter, we define a general model of evolution of sequences along a tree. Suppose that we have a κ-state ancestral sequence and a descendant sequence along a single edge of a tree. Note that the number of states κ for DNA is 4, and for proteins κ is 20. Choose one arbitrary site in the ancestral sequence. We need the probabilities that site may be occupied by each state. We also need the probabilities of base substitutions as this site in the ancestral sequence changes to a site in the descendant sequence over time. These model parameters can be summarized in a row vector and a matrix.

More formally, let T^ρ be a binary rooted topological tree with root ρ. Then the *general Markov model* of κ-state sequence evolution along T^ρ at a single site has the following parameters:

- A root distribution vector $\boldsymbol{\pi}$ which is a $1 \times \kappa$ vector whose entries are non-negative and add to 1. The entries are the probabilities of the various bases at a single site at the root ρ of the tree.

- A Markov matrix M^e for each edge e of the tree T^ρ directed away from the root. M^e is a $\kappa \times \kappa$ matrix whose entries are non-negative, with each row adding to 1. The i, j-entry of this matrix is the conditional probability of observing base j at the descendant node of e, if the base at the parent node is i.

For example, suppose that we have the tree shown in Figure 3.1 with root ρ and three leaves a, b, c representing extant species from which data can be obtained. Then there are four Markov matrices associated to the edges. Suppose that we have $\kappa = 2$ with state space $S = \{A, B\}$. Then for each edge, e, the Markov matrix is of the form:

$$M^e = \begin{pmatrix} p_{AA} & p_{AB} \\ p_{BA} & p_{BB} \end{pmatrix}.$$

The joint distribution P of site patterns at the leaves is a $2 \times 2 \times 2$ tensor where the entry $P(i, j, k)$ is the probability of observing i at a, j at b, and k at c where $i, j, k \in S$.

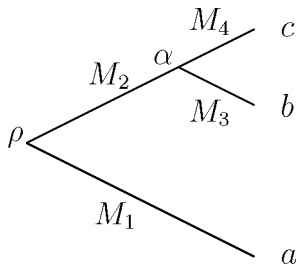

Figure 3.1: A 3-leaf rooted tree with general Markov parameters modeling the evolution of a site from a common ancestor ρ to 3 extant taxa, a, b, c. The internal node labeled α is the most recent common ancestor of b, c.

To compute $P(i, j, k)$ first suppose a single site has pattern A, A, B at the leaves a, b, c, that is $i = A, j = A, k = B$ respectively. Let $\boldsymbol{\pi} = (p_A, p_B)$ be the root distribution. To observe this pattern at the leaves, we might have A at ρ and A at the other internal node α, so that the probability for this particular case is:

$$p_A M_1(A, A) M_2(A, A) M_3(A, A) M_4(A, B).$$

Since states at internal nodes are not observable, as they occurred in the past, summing over all possibilities of bases at ρ and α gives $P(A, A, B)$:

$$\begin{aligned} P(A, A, B) = P(a = A, b = A, c = B) \\ = & p_A M_1(A, A) M_2(A, A) M_3(A, A) M_4(A, B) \\ + & p_A M_1(A, A) M_2(A, B) M_3(B, A) M_4(B, B) \\ + & p_B M_1(B, A) M_2(B, A) M_3(A, A) M_4(A, B) \\ + & p_B M_1(B, A) M_2(B, B) M_3(B, A) M_4(B, B). \end{aligned}$$

The seven other entries of P are obtained similarly.

We model evolution at each site in sequences as independent and identically distributed trials of the same process. Thus the joint distribution P describes the expected site frequency patterns at each site in the sequence, and entries of P could be estimated from data by using the observed frequencies of site patterns in aligned sequences.

For a rooted n-taxon tree and a κ-state character, the joint distribution, P, of bases at the leaves is a $\underbrace{\kappa \times \kappa \times \cdots \times \kappa}_{n}$ tensor. To compute any particular entry of P, by the same reasoning as for the 3-leaf tree, we sum over all possible states at the root and interior nodes, the product of entries in the matrices M^e for all edges e and $\boldsymbol{\pi}$ that result in a fixed site pattern $(i_1, \ldots, i_n)$ at the leaves of the tree.

In most model-based phylogenetic analyses, the Markov matrices come from a continuous-time model. A continuous-time model has a *rate matrix*, Q, which shows the instantaneous probabilistic rate of change between different bases. The off-diagonal entries of Q are non-negative, and the diagonals are set such that each row adds to 0. Now we can define the Markov matrix for an edge e in the tree of length $t_e > 0$ based on the rate matrix. Noting that the rate matrix can be rescaled such that the average rate of substitution is one, then the branch lengths of the phylogenetic tree can be considered in units of expected number of base changes per site. Now we can compute the associated Markov matrix on an edge of length t_e as $M^e = \exp(Qt_e)$.

Time-reversibility is one of the common features of Markov models used in practice. This means that the same model parameters can describe the evolution between the ancestral and descendant sequences on an edge regardless of the edge's orientation. In continuous-time models, time-reversibility holds if

$$\operatorname{diag}(\boldsymbol{\pi})Q = Q^T \operatorname{diag}(\boldsymbol{\pi}). \tag{3.1}$$

This implies that if P is the joint distribution of bases in the ancestral and descendant

sequences on a single edge, then $P = P^T$, and that $\boldsymbol{\pi}$ is a *stable base distribution* for M^e,

$$\boldsymbol{\pi} M^e = \boldsymbol{\pi}.$$

A time reversible model has the feature that the root of a tree might be relocated to any other point in the tree without making any other changes to the parameters, and the joint distribution of states at the leaves will be unchanged. This will be convenient in some of our proofs since it allows us to freely move the tree root around to simplify arguments.

The most common continuous-time model used for analyzing data is the *general time-reversible model* (GTR). Parameters in GTR models are $\boldsymbol{\pi}, Q$ satisfying equation 3.1, and a metric tree. The rate matrix, Q, for DNA is 4×4 and for amino acids is 20×20.

Equation 3.1 implies that for $i \neq j$,

$$q_{ij} = \pi_j r_{ij}, \tag{3.2}$$

for some $r_{ij} = r_{ji}$. The r_{ij} are called *relative rates* for the model and can be viewed together with the entries of $\boldsymbol{\pi}$ as independent parameters. This model was introduced into phylogenetics by Simon Tavaré [7] for DNA and Quang et al. [4] for proteins. The symmetric matrix $R = (r_{ij})$ is called the *exchangeability* matrix. Note that the diagonal entries of R are meaningless and may be set to 0 or something arbitrary.

Thus far we have assumed that every site in the sequences behaves identically in Markov models. Biologically, however, it is more reasonable to imagine classes of sites with different behavior which leads to a *mixture model*. For a finite number of classes and a fixed metric tree T with branch lengths $\{t_e\}$, there is a set of parameters for each class including the usual parameters $Q, \boldsymbol{\pi}$, for Markov models, and scaling parameters which are used to speed up and slow down the substitution process. Each class leads to a probability distribution of expected site frequency patterns and then we take a weighted sum of the probability distributions over the classes to describe the full model. For example, for two classes we

have a vector of weights, $(w, 1-w)$, $0 \leq w \leq 1$, which can be thought of as the relative sizes of the classes, and so the joint distribution of a mixture model is

$$P = wP_1 + (1-w)P_2.$$

In this equation, P_1 is the joint distribution for the first class and P_2 is the joint distribution for the second class.

The mixture model which this work focuses on is defined as the following:

Definition 1. *Let T be a rooted topological tree, κ the number of states and m the number of classes. Then the numerical parameters of the* Profile Mixture Model*, PM=PM (T, κ, m) are:*

- $\{t_e\}$, $t_e \geq 0$*; edge lengths of each edge e of T.*
- *R a symmetric $\kappa \times \kappa$ matrix of relative exchangabilities of states.*
- $\{\boldsymbol{\pi}_i\}_{i=1}^m$*; root distribution vectors of size $1 \times \kappa$ for each class i.*
- $\{r_i\}_{i=1}^m$, $r_i \geq 0$*; a scalar rate for each class i.*
- $\{w_i\}_{i=1}^m$*; class frequencies with $w_i > 0$ adding to 1.*

Note that the PM model uses a single matrix R of relative exchangeability for all classes, but that the $\boldsymbol{\pi}_i$ vary for the classes.

Definition 2. *For class i, the rate matrix is $Q_i = Q_i(R, \boldsymbol{\pi}_i)$, $\kappa \times \kappa$ with off-diagonal entries those of $R\,\mathrm{diag}(\boldsymbol{\pi}_i)$, and diagonal entries such that rows add to 0. Then the Markov matrix on edge e corresponding to class i is $M_i^e = \exp(Q_i t_e r_i)$.*

There is another way to think of mixture models as related to a κm-state model at internal nodes of the tree. Using notations as above, we have the following.

Definition 3. *Let Π_r be the κm-entry block vector of $\boldsymbol{\pi}_i$'s. Then $Q = Q(R, \Pi_r)$ is a block diagonal matrix of $r_i Q_i$'s, and the Markov matrix of edge e, $M^e = \exp(Qt_e)$, is a block diagonal matrix of the M_i^e's.*

Using M^e in Definition 3 and a κm-element root distribution of the form

$$(w_1\boldsymbol{\pi}_1, w_2\boldsymbol{\pi}_2, \ldots, w_\kappa\boldsymbol{\pi}_\kappa),$$

we obtain a tensor P describing the expected site pattern distribution with κm states at leaves for the tree T in which the classes have not been mixed. To mix the classes we can then multiply in each index of P by a $m\kappa \times \kappa$ matrix J of m stacked $\kappa \times \kappa$ identities I. Equivalently, we can instead replace the Markov matrices M^e on terminal edges by $M^e J$. The $M^e J$ have the following form.

Definition 4. *Let R, $\mathrm{diag}(\boldsymbol{\pi}_i)$, $r_i \geq 0$, $Q_i(R, \boldsymbol{\pi}_i)$, $M_i^e = \exp(Q_i t_e r_i)$, and Π_r be as in Definition 3, but $t_e \geq 0$ the length of terminal edge e. Then the Markov matrix on edge e, $M^e J$, is a $m\kappa \times \kappa$ stacked matrix of M_i^e's.*

As can be seen in Definition 1, the PM model has many parameters and it is crucial to know whether the parameters of the model can be recovered from the site pattern distribution they determine. This is the question of parameter *identifiability*. If a probability distribution comes from the model, can the parameters of the model which produced it, be recovered uniquely? In phylogenetics, we have two types of parameters: numerical parameters and the tree topology. Although we care most about identifying the tree and branch lengths, all parameters might be of biological interest and identifying the tree may depend on identifying the others as well. An old result that will be used in this work repeatedly is a basic identifiability result for the GTR model.

Theorem 1. *For the single class GTR model on an unrooted binary metric tree, all numerical parameters and the tree topology are generically identifiable.*

Generically identifiable means that all parameters are identifiable except those in a set of measure zero in the parameter space. The goal of this project is to obtain a similar result for the profile mixture model.

3.4 Algebraic Definitions and Lemmas

3.4.1 Definitions

In the preceding section, we have seen that the site pattern probabilities from a model form a $\kappa \times \kappa \times \cdots \times \kappa$ array or tensor. In this section we give some algebraic definitions and results concerning tensors that will be useful for studying such a model.

A key tool for this work is Kruskal's Theorem on the structure of certain 3-way tensors. To apply it, however, we will need to define an operation on matrices. Throughout this work, we let $[n] = \{1, 2, \ldots, n\}$.

Definition 5. *Let A be an $n \times l$ matrix and B be an $n \times m$ matrix. The* row tensor product *$A \otimes_r B$ is defined to be the $n \times lm$ matrix formed in the following way. Index columns by the ordered pairs (j, k), $j \in [\ell]$, $k \in [n]$, then*

$$(A \otimes_r B)_{i,(j,k)} = a_{ij} b_{ik}.$$

Note that the precise ordering of columns usually will not matter in this work since we are typically concerned with the rank of such a matrix. When an explicit ordering is needed, it will be made clear. There is also another more well-known type of tensor product of matrices which we define here.

Definition 6. *Let A be an $m \times k$ matrix and B be an $n \times l$ matrix. The* tensor product *$A \otimes B$ is defined to be $nm \times lk$ matrix whose rows are indexed by the ordered pair (i_1, j_1), $i_1 \in [m], j_1 \in [n]$ and whose columns are indexed by ordered pair (i_2, j_2), $i_2 \in [k], j_2 \in [l]$ such that*

$$(A \otimes B)_{(i_1,j_1),(i_2,j_2)} = a_{i_1 i_2} b_{j_1 j_2}.$$

The tensor product is also called the Kronecker product, and row and column indices may be given an explicit ordering if necessary.

In the case that all matrices are the same, we have the row tensor power.

Definition 7. *Let A be a $m \times n$ matrix and ℓ be an arbitrary positive integer. Then the ℓ^{th}* row-tensor power *of A is the $m \times n^\ell$ matrix*

$$A^{\otimes_r^\ell} = \underbrace{A \otimes_r A \otimes_r \cdots \otimes_r A}_{\ell}.$$

To apply Kruskal's Theorem we also need to define a flattening of a tensor with respect to some partition of its indices X, when $|X| = n$ for an n-way tensor. In the following definition, I and J are nonempty subsets of X, a set of n elements, such that

$$I \cap J = \emptyset, \qquad I \cup J = X.$$

They form a non-trivial bipartition or *split* of X, denoted by $I|J$.

Definition 8. *Let A be a $\underbrace{k \times k \cdots \times k}_{n}$ tensor with $I|J$ a split of the index set. Then the matrix flattening of A with respect to I, J, denoted* $\mathrm{Flat}_{I|J}(A)$*, is a $k^{|I|} \times k^{|J|}$ matrix. If by permuting indices, we assume that $I = \{1, 2, \cdots, |I|\}$, $J = \{|I| + 1, \cdots, n\}$, then*

$$\big(\mathrm{Flat}_{I|J}(A)\big)_{\alpha,\beta} = A(\alpha, \beta),$$

for $\alpha = (i_1, \ldots, i_{|I|}) \in k^{|I|}$ and $\beta = (j_1, \ldots, j_{|J|}) \in k^{|J|}$.

Similarly, we can make the definition of a tripartition and a flattening with respect to it. A *tripartition* of X here means three disjoint, nonempty sets whose union is X. For a tripartition of indices $I \sqcup J \sqcup K = X$, then assuming $I = \{1, 2, \cdots, |I|\}$, $J = \{|I| + 1, \cdots, |I| + |J|\}$, and $K = \{|I| + |J| + 1, \cdots, n\}$, for A of size $\underbrace{k \times k \cdots \times k}_{n}$, with indices $X = [n]$,

$$\left(\mathrm{Flat}_{I|J|K}(A)\right)_{\alpha,\beta,\gamma} = A(\alpha,\beta,\gamma),$$

is a 3-way tensor, where $\alpha \in k^{|I|}$, $\beta \in k^{|J|}$, and $\gamma \in k^{|K|}$.

Example 1. Let A be a $20 \times 20 \times 20 \times 20 \times 20 \times 20$ array. Let $\kappa = 20$, $I = \{1,3\}$, $J = \{4\}$, and $K = \{2,5,6\}$. Then $\mathrm{Flat}_{I|J|K}(A)$ is a $20^2 \times 20 \times 20^3$ tensor with, for example,

$$\left(\mathrm{Flat}_{I|J|K}(A)\right)_{(10,12),(8),(15,16,18)} = A(10,15,12,8,16,18).$$

Definition 9. *Let A be a $m \times n_A$ matrix with ith row $r_i^A = (r_i^A(1), \cdots, r_i^A(n_A))$. Similarly B, and C with m rows, and n_B, n_C columns respectively. Then $[A,B,C]$ denotes the 3-way $n_A \times n_B \times n_C$ tensor*

$$[A,B,C] = \sum_{i=1}^{m} r_i^A \otimes r_i^B \otimes r_i^C.$$

The following example illustrates Definition 9:

Example 2. Let A, B, C be 2×2, 2×3, and 2×4 matrices respectively,

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix},$$

then $P = [A,B,C]$ is a $2 \times 3 \times 4$ tensor as following:

$$P(:,:,1) = \begin{pmatrix} 61 & 77 & 93 \\ 82 & 104 & 126 \end{pmatrix} \quad P(:,:,2) = \begin{pmatrix} 74 & 94 & 114 \\ 100 & 128 & 156 \end{pmatrix}$$
$$P(:,:,3) = \begin{pmatrix} 87 & 111 & 135 \\ 118 & 152 & 186 \end{pmatrix} \quad P(:,:,4) = \begin{pmatrix} 100 & 128 & 156 \\ 136 & 176 & 216 \end{pmatrix}.$$

Also we can extend Definition 9 slightly to define a 3-way tensor $[\boldsymbol{\pi}; A, B, C]$ where

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_m)$$

is a $m \times 1$ vector:

$$[\boldsymbol{\pi}_i; A, B, C] = \sum_{j=1}^{m} \pi_i r_j^A \otimes r_j^B \otimes r_j^C.$$

Before stating Kruskal's Theorem, we need the notion of the Kruskal row rank of a matrix.

Definition 10. *Let A be a matrix. The* Kruskal (row) rank *of A is the largest number k such that every set of k rows of A are independent.*

Note that the usual definition of rank of a matrix is the largest number k such that some set of k rows are independent. Thus the Kruskal rank may be less than the usual rank.

Theorem 2 (Kruskal [3])**.** *Let A, B, C be $l \times n_A$, $l \times n_B$, and $l \times n_C$ matrices with Kruskal rank p, q, r respectively. If*

$$p + q + r \geq 2l + 2, \tag{3.3}$$

then A, B, C are uniquely determined by $[A, B, C]$ up to simultaneous permutation and scaling of the rows. More precisely, if $[A, B, C] = [A', B', C']$ then there exist invertable diagonal matrices D_1, D_2 and a permutation P such that

$$A' = PD_1A, \quad B' = PD_2B, \quad C' = PD_1^{-1}D_2^{-1}C.$$

Note that for two matrices A, B, the natural extension of the bracket notation gives $[A, B] = A^TB$. So from $[A, B]$, A and B cannot be determined uniquely, since there are many matrix products that give the same product, for instance $A^TB = (QA)^T(QB)$ for any orthogonal matrix Q.

Thus Kruskal's theorem shows a significant difference between matrices and 3-way tensors. The following example illustrates the issue with permutation and scaling for the 3-way bracket product.

Example 3. Let A, B, C be as in Example 2, and $A_1 = P(2A), B_1 = P(3B)$, and $C_1 = P\frac{1}{6}C$ where P interchanges the row order.

$$A_1 = \begin{pmatrix} 6 & 8 \\ 2 & 4 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 12 & 15 & 18 \\ 3 & 6 & 9 \end{pmatrix},$$
$$C_1 = \begin{pmatrix} \frac{5}{6} & \frac{6}{6} & \frac{7}{6} & \frac{8}{6} \\ \frac{1}{6} & \frac{2}{6} & \frac{3}{6} & \frac{4}{6} \end{pmatrix},$$

then $P_1 = [A_1, B_1, C_1]$ is a $2 \times 3 \times 4$ tensor:

$$P_1(:,:,1) = \begin{pmatrix} 61 & 77 & 93 \\ 82 & 104 & 126 \end{pmatrix} \quad P_1(:,:,2) = \begin{pmatrix} 74 & 94 & 114 \\ 100 & 128 & 156 \end{pmatrix}$$
$$P_1(:,:,3) = \begin{pmatrix} 87 & 111 & 135 \\ 118 & 152 & 186 \end{pmatrix} \quad P_1(:,:,4) = \begin{pmatrix} 100 & 128 & 156 \\ 136 & 176 & 216 \end{pmatrix}.$$

So $[A, B, C] = [A_1, B_1, C_1]$.

In order to apply Kruskal's theorem for matrices coming from biological models, we may face some cases which are exceptional and do not have the Kruskal rank needed for an argument. The best way to talk about these exceptional cases is using the language of algebraic varieties. By an example, we will get better understanding.

Example 4. Let M be the set of all 3×3 matrices. Then the dimension of this set is 9. The matrices with Kruskal rank 0 in this set are, up to ordering of rows:

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} a & b & c \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 0 \end{pmatrix},$$

which have a row of zeros. Up to ordering and independent non-zero rescaling of the rows,

the matrices with Kruskal rank 1 are:

$$\begin{pmatrix} a & b & c \\ a & b & c \\ a & b & c \end{pmatrix}, \quad \begin{pmatrix} a & b & c \\ a & b & c \\ d & e & f \end{pmatrix},$$

where $(a,b,c) \neq (0,0,0), (d,e,f) \neq (0,0,0)$, which have all non-zero rows, with two that are multiples of one another. Those with Kruskal rank 2 are, up to ordering:

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix},$$

where all rows are non-zero, (a,b,c) and (d,e,f) are independent, and $(g,h,i) = \alpha(a,b,c) + \beta(d,e,f)$ with $\alpha, \beta \neq 0$.

Notice that all of the sets of matrices of these particular forms have lower dimension inside 9-dimensional space (dimension at most 8). Then in that sense, most of the matrices in M have Kruskal rank 3.

In order to make this more precise, we define the notation of a variety.

Definition 11. *Let S be a finite set of polynomials in $\mathbb{C}[a_1, \ldots, a_n]$. The zero set in $\mathbb{C}^n$ of polynomials in S is called the* algebraic variety *$V(S)$ associated to S. Similarly, the zero set in $\mathbb{C}^n$ of a finite set of analytic functions is called an analytic variety. A subset of a variety that is itself a variety is called a* subvariety*.*

Two different sets S, S' can have $V(S) = V(S')$, so we define a largest set of polynomials defining a variety, called an *ideal*.

Definition 12. *Given an algebraic or analytic variety $V(S)$, the ideal $I(V(S))$ is the set of all polynomials $p \in \mathbb{C}[a_1, \ldots, a_n]$ or analytic functions in the a_i such that $p(v) = 0$ for all $v \in V(S)$. Thus $S \subset I(V(S))$.*

Definition 13. *A variety V is said to be* irreducible *if it cannot be expressed as $V = V_1 \cup V_2$ where V_1, V_2 are varieties with $V_1, V_2 \subsetneq V$.*

The main proposition of this section will be used to show model parameters are identifiable except for "rare" choices. In an algebraic setting, this is expressed using the following terminology.

Definition 14. *A property is* generic *on a variety V if it holds for all points on the variety V except possibly for those points in some proper subvariety $U \subsetneq V$. When a property is generic on V, we will also say the property holds for generic points on V. If V is irreducible, by principles of algebraic geometry, U must be of lower dimension than V, and thus of measure 0 in V.*

Example 5. The property of having Kruskal rank 3 holds for generic 3×3 matrices. To see this, observe in Example 4 that all matrices of Kruskal rank 2 or less have rank 2 or less, so they lie in the subvariety defined by the 3×3 determinants. This also shows the property of having full rank is generic for 3×3 matrices.

Our main use of varieties in this project is through their connection to generic identifiability. The following proposition provides the means of drawing generic conclusions.

Proposition 15. *[5] Let V_0 and V_1 be two varieties, with V_1 irreducible. Suppose $f_0 \in I(V_0)$, and there exists a point $p_1 \in V_1$ with $f_0(p_1) \neq 0$. Then $V_1 \nsubseteq V_0$, and the variety $V_0 \cap V_1 = V(I(V_0) \cup I(V_1))$ is of lower dimension than V_1. That is, generic points on V_1 lie off of V_0.*

We also need the following.

Proposition 16. *Suppose V is the smallest variety containing the image of a polynomial or analytic parametrization. Then V is irreducible, of the same dimension as the image of the parametrization.*

The smallest variety containing a set is called the *Zariski Closure* of the set.

3.4.2 Rank Propositions

Since we will need to know the Kruskal rank of Markov matrices, and some related ranks, coming from the PM model defined in Section 3.3, here we give some computational results using `PARI/GP` about the row rank and the Kruskal row rank of such matrices. In Section 3.6, when we apply Proposition 15, the point p_1 will be chosen to have one of these forms.

Definition 17. *Let $a_1, \ldots, a_\kappa$ be arbitrary complex numbers. Then $M(a_1, \ldots, a_\kappa)$ denotes a $\kappa \times \kappa$ matrix of the form*

$$M(a_1, \ldots, a_\kappa) = \begin{pmatrix} 1 + a_1 - s & a_2 & \cdots & a_\kappa \\ a_1 & 1 + a_2 - s & \cdots & a_\kappa \\ \vdots & \ddots & & \vdots \\ a_1 & a_2 & \cdots & 1 + a_\kappa - s \end{pmatrix}$$

where $s = a_1 + \cdots + a_\kappa$.

Proposition 18. *Let M be a $m\kappa \times \kappa$ matrix formed by stacking m distinct matrices of form $M(a_1, \ldots, a_\kappa)$. Let $\kappa = 20$ and $m \leq 77$. Then $M^{\otimes_r^\ell}$ has full row rank for $\ell \geq 3$ and generic entries of M.*

Proof. We begin with the special case of $\ell = 3$. An exact `PARI/GP` calculation (see code in Appendix) shows that by picking distinct random integers for $a_1, \ldots, a_\kappa$ in each block in M, with $\kappa = 20$ and $m = 77$, we may find some $p_0 = M$ for which $M^{\otimes_r^\ell}$ has full row rank when $\ell = 3$. This implies the same for $m \leq 77$.

The set of all such M is defined by linear polynomials on $\mathbb{C}^{m\kappa^2}$ and thus is a variety V_1. Since V_1 is parameterized, it is irreducible. Let f_0 be any $m\kappa \times m\kappa$ minor of $M^{\otimes_r^\ell}$, as a polynomial in the entries of M, such that $f_0(p_0) \neq 0$. Then by Proposition 15 the set of M for which $f_0(M) = 0$ is a lower dimension subvariety of V_1. That is, generic M give $M^{\otimes_r^\ell}$ of rank $m\kappa$.

Now consider $\ell \geq 3$. Then

$$M^{\otimes_r^\ell} = M^{\otimes_r^3} \otimes_r M^{\otimes_r^{\ell-3}},$$

where $M^{\otimes_r^3}$ is a $m\kappa \times q$ matrix and $M^{\otimes_r^{\ell-3}}$ is a $m\kappa \times u$ matrix. Since $M^{\otimes_r^3}$ has full row rank $m\kappa$ for generic M, its rows are independent. Since

$$M^{\otimes_r^3} \otimes_r M^{\otimes_r^{\ell-3}} = \begin{pmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1q} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2q} \\ \vdots & \ddots & \vdots & \\ \mu_{p1} & \mu_{p2} & \cdots & \mu_{pq} \end{pmatrix} \otimes_r \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1u} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2u} \\ \vdots & \ddots & \vdots & \\ \alpha_{p1} & \alpha_{p2} & \cdots & \alpha_{pu} \end{pmatrix}$$

$$= \begin{pmatrix} \mu_{11}\alpha_{11} & \mu_{12}\alpha_{11} & \cdots & \mu_{1q}\alpha_{11} & \cdots \\ \mu_{21}\alpha_{21} & \mu_{22}\alpha_{21} & \cdots & \mu_{2q}\alpha_{21} & \cdots \\ \vdots & \ddots & \vdots & & \\ \mu_{p1}\alpha_{p1} & \mu_{p2}\alpha_{p1} & \cdots & \mu_{pq}\alpha_{p1} & \cdots \end{pmatrix},$$

it is enough to know the entries of the some single column of $M^{\otimes_r^{\ell-3}}$ are nonzero to ensure $M^{\otimes_r^\ell}$ has $m\kappa$ independent rows. This is true for generic choice of M. □

Remark 19. *Note that for $\ell = 2$, $M^{\otimes_r^\ell}$ does not appear to generically have full row rank, as* `PARI/GP` *calculations show. Also it has been observed that when you increase the number m of classes by one, the difference between the number of rows and the rank of $M^{\otimes_r^\ell}$ follows the Fibonacci sequence until the rank of $M^{\otimes_r^\ell}$ hits the number of independent columns.*

The next proposition in this section is a result about the Kruskal row rank, valid for all $M^{\otimes_r^\ell}$.

Proposition 20. *For $\kappa \geq 2$, let M be a $m\kappa \times \kappa$ matrix formed by stacking m distinct matrices of form $M(a_1, \ldots, a_\kappa)$. For $\ell \geq 1$, $M^{\otimes_r^\ell}$ has Kruskal row rank greater than or equal to 2 for generic entries of the M.*

Proof. Consider first the case $\ell = 1$. The matrices of Kruskal row rank at most one lie inside the variety of matrices row rank of at most 1, is defined by the ideal generically by all 2×2 minors. To see that generic matrices of the given form have Kruskal row rank 2 or greater, by Proposition 15 it is enough to find one such matrix not on this variety. Choose $m\kappa$ distinct positive small numbers as the free entries $a_1, \ldots, a_\kappa$ in each block of in M. Since each block of M is in the form of $M(a_1, \ldots, a_\kappa)$, then no two rows of $M(a_1, \ldots, a_\kappa)$ are multiples of each other, since the diagonal entries are the largest. No two rows of different blocks are multiples, because the a_i's are distinct. Thus M has rank greater than or equal to two.

The cases $\ell > 1$ now follows by an argument similar to that at the end of the proof of Proposition 18. □

Proposition 21. *Let M be a $m\kappa^2 \times \kappa^3$ matrix formed by stacking m matrices of the form $M(a_1, \cdots, a_\kappa)^{\otimes_r^2} \otimes M(a_1, \cdots, a_\kappa)$. Then for $\kappa = 20$ and $m < 77$, a generic M has rank $> m\kappa$.*

Proof. A `PARI/GP` calculation shows that for random integer values of a_i's, M has

- full row rank 400 when $m = 1$, which is greater than $m\kappa = 20$,
- full row rank 800 when $m = 2$, which is greater than $m\kappa = 40$,
- rank 1180 when $m = 3$, which is greater than $m\kappa = 60$,
- rank 1540 when $m = 4$, which is greater than $m\kappa = 80$.

For $m = 5, 6, \cdots, 77$, the rank is at least $1540 = 20 \times 77$ for some choice of a_i's, which shows that for $\kappa = 20$ and $m < 77$, M has rank $> m\kappa$.

The generic rank in all these cases must be at least as large as for these random examples.

□

Proposition 22. *Let M_1 be of the form M in Proposition 21, and M_2 be formed by stacking matrices of the form $M(a_1, \cdots, a_\kappa) \otimes M(a_1, \cdots, a_\kappa)^{\otimes_r^2}$. Let L be a $m\kappa^2 \times m\kappa^2$ diagonal matrix with positive entries. Then for $\kappa = 20$ and $m < 74$, $M_2^T L M_1$ has rank $> m\kappa$.*

Proof. Sylvester's rank inequality gives

$$\operatorname{rank}(M_2^T L M_1) \geq \operatorname{rank}(LM_1) + \operatorname{rank}(M_2^T) - m\kappa^2.$$

Since L is a diagonal matrix with positive entries, then the rank of LM_1 is the same as the rank of M_1. Then by calculations on the proof of Proposition 21 there is all choice of a_i's so that $M_2^T L M_1$ has rank at least

- $400 + 400 - 400 = 400$ when $m = 1$, which is greater than $m\kappa = 20$,
- $800 + 800 - 800 = 800$ when $m = 2$, which is greater than $m\kappa = 40$,
- $1180 + 1180 - 1200 = 1160$ when $m = 3$, which is greater than $m\kappa = 60$,
- $1540 + 1540 - 1600 = 1480$ when $m = 4$, which is greater than $m\kappa = 80$.

Since this choice of a_i's for $m = 4$ can be extended to larger m by repeating blocks for the new classes, this shows there are choices of a_i's giving $M_2^T L M_1$ rank at least 1480 for all $m \geq 4$. Finally, $1480 > 20m$ for $4 \leq m < 74$. These examples imply the generic rank must be $> m\kappa$. □

3.5 Algebraic Aspects of the Profile Mixture Model

In this section we relate some of the algebraic definitions we made in the previous section to the PM model. We begin by describing how a row tensor product of Markov matrices relates to a star tree.

Definition 23. *Let A be a set of taxa on a star tree rooted at its internal node with pendant edges $\{e_1, \ldots, e_{|A|}\}$ and associated Markov matrices M^{e_i}. Then*

$$M_A = M^{e_1} \otimes_r \cdots \otimes_r M^{e_{|A|}}$$

where $\otimes_r$ denotes the row tensor product.

The entries of M_A represent the conditional probability of observing different states at taxa in set A given the state at the root. This is because the entries of each M^{e_i} represent conditional probabilities of observing states at one taxon and phylogenetic models assume independence of substitution processes on different edges. Since each M^{e_i} in this equation by Definition 4 is a $m\kappa \times \kappa$ matrix, then M_A is a $m\kappa \times \kappa^{|A|}$.

Next we apply the concept of flattening to a probability distribution for a Markov model on a tree. But since the flattening is defined with respect to some tripartition of a set of taxa, we need to explain what it means for a tripartition to be displayed on a tree. Let $A = \{a, b, c\}, B = \{d, f\}, C = \{g, h\}$ be three disjoint subsets of a set of taxa $X = \{a, b, c, d, f, g, h\}$. Then the tree of Figure 3.2 displays this tripartition of X. More formally, a tripartition $A|B|C$ is displayed on a tree if there is some vertex v whose deletion results in three subtrees with the elements of the sets A, B, C labeling the leaves on each.

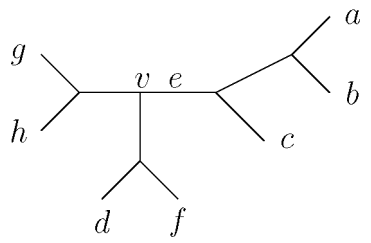

Figure 3.2: A tree displaying $A|B|C$ with $A = \{a, b, c\}, B = \{d, f\}, C = \{g, h\}$ since deletion of the vertex v partitions the leaves into these sets.

Similarly we can have a bipartition or split displayed on a tree. For example in Figure 3.2, if $A = \{a, b, c\}$ and $B = \{d, f, g, h\}$, then the tree displays the bipartition $A|B$ of the set of taxa X. More formally, a split $A|B$ is displayed on a tree if there is an edge e whose deletion results in two subsets with leaves labeled by the elements of A and B.

Now we claim that if we have a tree T displaying a tripartition of a set of taxa, then the flattening of a joint distribution of bases at the leaves from a Markov model can be expressed using the 3-way matrix product defined in Section 3.4.

Lemma 24. *Suppose T is a tree on a set of taxa X displaying a tripartition $A|B|C$. Let*

P be a probability distribution for a Markov model on T with s states at the internal nodes. Then for some matrices $\tilde{M}_A, \tilde{M}_B, \tilde{M}_C$, each with s rows,

$$\mathrm{Flat}_{A|B|C}(P) = [\tilde{M}_A, \tilde{M}_B, \tilde{M}_C].$$

Proof. Suppose T has a known tripartition $A|B|C$ at vertex v. Now we can define matrices M_A, M_B, M_C whose entries are conditional probabilities of states at the leaves in each set A, B, C, given the state at v, from the parameters on T. Let $\boldsymbol{\pi}$ be the base distribution at v. Then

$$\mathrm{Flat}_{A|B|C}(P) = [\boldsymbol{\pi}; M_A, M_B, M_C].$$

Let $\overline{M}_A$ be a matrix whose rows are obtained by the product of M_A's rows and the corresponding entry of $\boldsymbol{\pi}$, i.e. $\overline{M}_A = \mathrm{diag}(\boldsymbol{\pi})M_A$. Then

$$\mathrm{Flat}_{A|B|C}(P) = [\overline{M}_A, M_B, M_C].$$

Taking $\tilde{M}_A = \overline{M}_A$, $\tilde{M}_B = M_B$, and $\tilde{M}_C = M_C$ gives us the result. □

In working with the PM model, so we will need to know the form of the Markov matrices when we fix the exchangeability matrix R. In the following lemma we prove that for a special R, a Markov matrix for one class has the particular form given in Definition 17.

Lemma 25. *Suppose that the parameter R of the PM model is a $\kappa \times \kappa$ matrix whose entries are all 1. Let e be an edge with $t_e = 1$. Then the Markov matrix for class i, M_i^e, is the form of*

$$M(a_1, \ldots, a_\kappa) = \begin{pmatrix} 1 + a_1 - s & a_2 & \cdots & a_\kappa \\ a_1 & 1 + a_2 - s & \cdots & a_\kappa \\ \vdots & \ddots & & \vdots \\ a_1 & a_2 & \cdots & 1 + a_\kappa - s \end{pmatrix}.$$

Proof. Suppose that R is a $\kappa \times \kappa$ matrix whose entries are all 1.

$$R = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \ddots & \vdots & \\ 1 & 1 & \cdots & 1 \end{pmatrix}.$$

For a class i with the rate r_i, let $\boldsymbol{\pi}_i = (\pi_1, \cdots, \pi_\kappa)$ whose non-negative entries sum to 1. Then

$$R\ (\text{diag}\ \boldsymbol{\pi}_i) = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \ddots & \vdots & \\ 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ \vdots & \ddots & \vdots & \\ 0 & 0 & \cdots & \pi_\kappa \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_\kappa \\ \pi_1 & \pi_2 & \cdots & \pi_\kappa \\ \vdots & \ddots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_\kappa \end{pmatrix},$$

and since $\sum_{j=1}^{\kappa} \pi_j = 1$, then the rate matrix is

$$Q_i = \begin{pmatrix} \pi_1 - 1 & \pi_2 & \cdots & \pi_\kappa \\ \pi_1 & \pi_2 - 1 & \cdots & \pi_\kappa \\ \vdots & \ddots & \vdots & \\ \pi_1 & \pi_2 & \cdots & \pi_\kappa - 1 \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_\kappa \\ \pi_1 & \pi_2 & \cdots & \pi_\kappa \\ \vdots & \ddots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_\kappa \end{pmatrix} - I = R(\text{diag}\ \boldsymbol{\pi}_i) - I.$$

By finding eigenvectors and eigenvalues for Q_i and using the diagonalization formula we have

$$Q_i = S_i \Lambda S_i^{-1},$$

with

$$\Lambda = \text{diag}\,(0, -1, -1, \cdots, -1),$$

$$S = \begin{pmatrix} 1 & -\frac{\pi_2}{\pi_1} & -\frac{\pi_3}{\pi_1} & \cdots & -\frac{\pi_\kappa}{\pi_1} \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix}, \quad S^{-1} = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 & \cdots & \pi_\kappa \\ -\pi_1 & 1-\pi_2 & -\pi_3 & \cdots & -\pi_\kappa \\ -\pi_1 & -\pi_2 & 1-\pi_3 & \cdots & -\pi_\kappa \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\pi_1 & -\pi_2 & -\pi_3 & \cdots & 1-\pi_\kappa \end{pmatrix}.$$

Then the Markov matrix for an edge of length $t_e = 1$ is

$$M_i^e = e^{Q_i r_i} = S e^{\Lambda r_i} S^{-1} = S \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & e^{-r_i} & \cdots & 0 \\ \vdots & \ddots & \vdots & \\ 0 & 0 & \cdots & e^{-r_i} \end{pmatrix} S^{-1}$$

$$= \begin{pmatrix} 1+a_1-s & a_2 & \cdots & a_\kappa \\ a_1 & 1+a_2-s & \cdots & a_\kappa \\ \vdots & \ddots & & \vdots \\ a_1 & a_2 & \cdots & 1+a_\kappa-s \end{pmatrix}$$

where $a_j = \pi_j(1-e^{r_i}) \geq 0$ for $j = 1, \cdots, \kappa$, and $s = a_1 + a_2 + \cdots + a_\kappa$. □

Now we show conversely that if we have a Markov matrix of the form of $M(a_1, \ldots, a_\kappa)$, it comes from an exponential of a rate matrix.

Lemma 26. *Given a Markov matrix, M_i^e of the form of $M(a_1, \ldots, a_\kappa)$ with $a_j \geq 0$ and $s = \sum_{j=1}^{\kappa} a_j$ suppose $0 < s \leq 1$. Then there is a $\boldsymbol{\pi}_i$ and r_i such that for R the matrix of all 1s and Q_i given by Definition 2, $M_i^e = \exp(r_i Q_i)$.*

Proof. Since $0 < s \leq 1$, there is a $r_i > 0$ such that $s = 1 - e^{-r_i}$. Let $\pi_j = \frac{a_j}{s}$ for $j = 1, \cdots, \kappa$.

Then $\sum_{j=1}^{\kappa} \pi_j = 1$, and $a_j = \pi_j(1 - e^{-r_i})$. Take

$$Q_i = \begin{pmatrix} \pi_1 - 1 & \pi_2 & \cdots & \pi_\kappa \\ \pi_1 & \pi_2 - 1 & \cdots & \pi_\kappa \\ \vdots & \ddots & \vdots & \\ \pi_1 & \pi_2 & \cdots & \pi_\kappa - 1 \end{pmatrix}.$$

Then $M_i^e = \exp(r_i Q_i)$. □

3.6 Identifiability of Parameters for the Profile Mixture Model

In this section, we prove the main result that the tree parameter and numerical parameters of the PM model are generically identifiable. We begin by investigating the rank of the flattening of a probability distribution array with respect to a bipartition of taxa.

Proposition 27. *Let T be an n-taxon tree on X and consider a distribution P from the model PM=PM $(T, 20, m)$ with $m < 77$. Suppose that $A|B$ is a split of X with $|A|, |B| \geq 3$.*

(1) If $A|B$ is displayed on T, then $\mathrm{Flat}_{A|B}(P)$ has rank less than or equal to $m\kappa$;

(2) If $A|B$ is not displayed on T, then $\mathrm{Flat}_{A|B}(P)$ generically has rank greater than $m\kappa$.

Here "generically" means for all choice of numerical parameters except those in a subset of measure zero.

Before proving this result, an example gives a better understanding of the form of $\mathrm{Flat}_{A|B}(P)$.

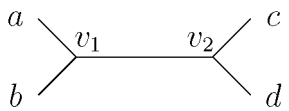

Figure 3.3: A 4-taxon tree with $\{a, b\}|\{c, d\}$ split.

Let $\kappa = 2$, with states denoted $1, 2$, and $m = 1$. Let $A = \{a, b\}$ and $B = \{c, d\}$ so that $\{a, b\}|\{c, d\}$ is a split on the tree T shown in Figure 3.3. Then P is the $2 \times 2 \times 2 \times 2$

joint probability tensor of observations at the leaves of the tree, so $P(i,j,k,l) = p_{ijkl}$ is the probability of observing state i at a, j at b, k at c, and l at d. Then the matrix $\text{Flat}_{A|B}(P)$ is

$$\text{Flat}_{A|B}(P) = \begin{array}{c} \\ (1,1) \\ (1,2) \\ (2,1) \\ (2,2) \end{array} \begin{array}{c} \begin{array}{cccc} (1,1) & (1,2) & (2,1) & (2,2) \end{array} \\ \begin{pmatrix} p_{1111} & p_{1112} & p_{1121} & p_{1122} \\ p_{1211} & p_{1212} & p_{1221} & p_{1222} \\ p_{2111} & p_{2112} & p_{2121} & p_{2122} \\ p_{2211} & p_{2212} & p_{2221} & p_{2222} \end{pmatrix} \end{array}$$

where the rows correspond to possible states at a and b, and the columns to possible states at c and d.

Consider the case where the terminal edges of the tree have length 0, so no substitutions occur on them. Then since $A|B$ is displayed on T, there are many zeros in $\text{Flat}_{A|B}(P)$, because the states at a and b must agree, as must those at c and d, for an entry to be non-zero.

Then $\text{Flat}_{A|B}(P)$ essentially just describes the joint distribution of states at v_1 and v_2 and has form

$$\text{Flat}_{A|B}(P) = \begin{array}{c} \\ (1,1) \\ (1,2) \\ (2,1) \\ (2,2) \end{array} \begin{array}{c} \begin{array}{cccc} (1,1) & (1,2) & (2,1) & (2,2) \end{array} \\ \begin{pmatrix} p_{1111} & 0 & 0 & p_{1122} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ p_{2211} & 0 & 0 & p_{2222} \end{pmatrix} \end{array},$$

which has rank at most $2 = m\kappa$.

Now let $A' = \{a, c\}, B' = \{b, d\}$. Since $A'|B'$ is not displayed on T, then with taxa ordered as a, b, c, d,

$$\mathrm{Flat}_{A'|B'}(P) = \begin{array}{c} \\ (1,1) \\ (1,2) \\ (2,1) \\ (2,2) \end{array} \begin{array}{c} \begin{array}{cccc} (1,1) & (1,2) & (2,1) & (2,2) \end{array} \\ \begin{pmatrix} p_{1111} & 0 & 0 & 0 \\ 0 & p_{1122} & 0 & 0 \\ 0 & 0 & p_{2211} & 0 \\ 0 & 0 & 0 & p_{2222} \end{pmatrix} \end{array}.$$

This matrix generically has rank 4, that is, $(m\kappa)^2$. Now we prove Proposition 27.

Proof of Proposition 27. To show claim (1), suppose the split $A|B$ is displayed on T and has associated edge $e = (\alpha, \beta)$. Conditioning on the state at α, let M_A be the matrix of size $m\kappa \times \kappa^{|A|}$ whose jth row entries $(j, \boldsymbol{k})$, for $\boldsymbol{k} \in [\kappa]^{|A|}$, are the probabilities of jointly observing the states $\boldsymbol{k}$ on A conditioned on state j at α. Similarly, let M_B be a $m\kappa \times \kappa^{|B|}$ matrix that describes the probability of jointly observing the states on B conditioned on the state at α. Then by rooting the tree at β, since the joint distribution at α, β is $\mathrm{diag}(\boldsymbol{\pi})M^e$ it follows that

$$\mathrm{Flat}_{A|B}(P) = M_A^T \,\mathrm{diag}(\boldsymbol{\pi})M^e M_B.$$

Since the Markov matrix M^e associated to e is $m\kappa \times m\kappa$, this factorization shows that $\mathrm{Flat}_{A|B}(P)$ has rank at most $m\kappa$.

Now suppose $A|B$ is not displayed on T, in order to show claim (2) that the rank of $\mathrm{Flat}_{A|B}(P)$ is generically greater than $m\kappa$. Let V_0 be the variety of matrices of the same size as $\mathrm{Flat}_{A|B}(P)$ with rank at most $m\kappa$, defined by the $m\kappa \times m\kappa$ minors. Let V_1 be the Zariski closure of the image of the parametrization. Then finding a single choice of parameters producing a rank greater than $m\kappa$ gives a point on $V_1 \setminus V_0$. This will yield the result by Propositions 15 and 16.

Since T does not display $A|B$, by Theorem 3.8.6 of Semple and Steel [6], there is an edge $e = (v_1, v_2)$ of T with associated split $C|D$ such that $A' = A \cap C$, $A'' = A \cap D$, $B' = B \cap C$, $B'' = B \cap D$ are all nonempty. To find the needed choice of parameters, first fix all internal

edges of T except e to have length 0, so the Markov matrices of these edges are I. Fix the edge length for all terminal edges and for e to be 1. Then choose all entries of R to be 1.

We have not yet specified value for the parameters $\boldsymbol{\pi}_i, r_i$ as these will be given values later in the argument. We are effectively reducing to the case of T being formed by two star trees, one on C and one on D, all of whose branches are of equal length, connected by edge e. See Figure 3.4. In addition, we can "sort" the edges to C and D into subgroups A', B', A'', B'' as shown.

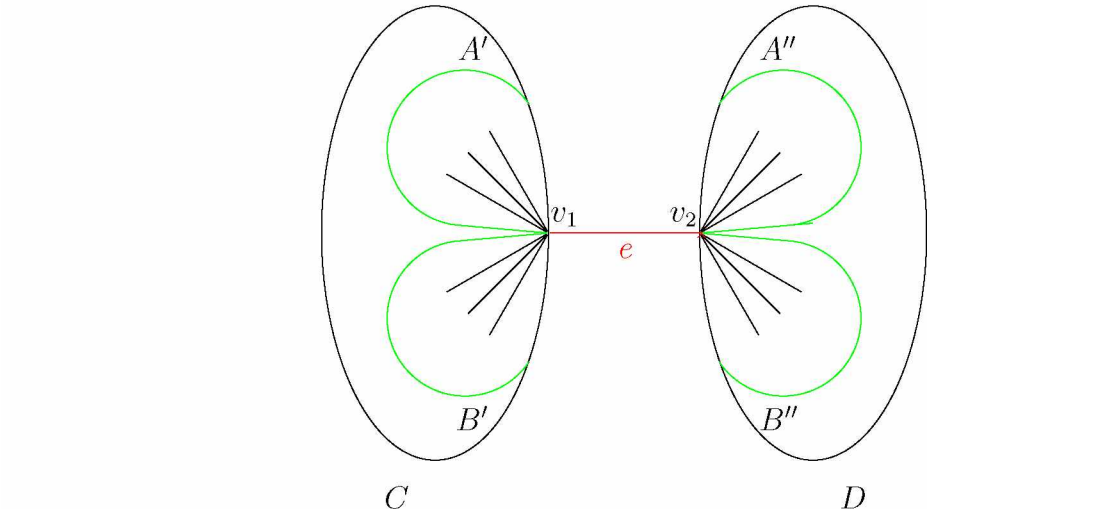

Figure 3.4: A tree which does not display the split $A|B$, but displays the split $C|D$ such that $A' = A \cap C$, $A'' = A \cap D$, $B' = B \cap C$, $B'' = B \cap D$.

Denote one of the endpoints of e as a root, say $r = v_1$. Let $K = \mathrm{diag}(\boldsymbol{\pi}_r)M^e$ be the $m\kappa \times m\kappa$ block diagonal matrix which expresses the joint distributions of states at the vertices v_1 and v_2 of e. The probabilities of observing states $\boldsymbol{i}$, $\boldsymbol{j}$, $\boldsymbol{k}$, $\boldsymbol{l}$ at leaves in A', B', A'', B'' respectively, $P(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}, \boldsymbol{l})$, are the entries of a $\kappa^{|A'|} \times \kappa^{|B'|} \times \kappa^{|A''|} \times \kappa^{|B''|}$ tensor. For understanding the structure of the flattening, we focus on the central edge e. We define a $m\kappa \times m\kappa \times m\kappa \times m\kappa$ tensor with $(m\kappa)^2$ non-zero entries,

$$Q(i,j,k,l) = \begin{cases} K(i,k) & i = j, k = l, \\ 0 & \text{otherwise}, \end{cases}$$

where K expresses the joint distribution of states at the vertices v_1 and v_2. The tensor Q represents the joint distribution of the tree of Figure 3.4 if terminal edges have length zero and A', B', A'', B'' have single taxa. Since $A|B$ is not displayed on T, in this case $\text{Flat}_{A|B}(Q)$ is a $(m\kappa)^2 \times (m\kappa)^2$ matrix $\hat{Q}$ whose entries are

$$\text{Flat}_{A|B}(Q) = \hat{Q}((i,j),(k,l)) = Q(i,k,j,l),$$

which is diagonal and generically of rank $m\kappa^2$.

Since P "expands" these terminal edges to edges with $m\kappa \times \kappa$ Markov matrices, it is plausible that the rank of the flattening of P is also large. Let $N_A = M_{A'} \otimes M_{A''}$ and $N_B = M_{B'} \otimes M_{B''}$ where $M_{A'}, M_{A''}, M_{B'}, M_{B''}$ are defined as before. Then

$$\text{Flat}_{A|B}(P) = N_A^T \hat{Q} N_B. \tag{3.4}$$

First we prove that claim (2) is true when $|A| = |B| = 3$ by considering two cases. Suppose first that $|A'| = |B'| = 2$ and $|A''| = |B''| = 1$, as shown in Figure 3.5.

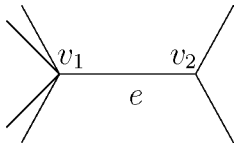

Figure 3.5: A tree such as in Figure 3.4 with $|A'| = |B'| = 2$ and $|A''| = |B''| = 1$.

Note that the probability of states i, k at v_1, v_2 is zero if i, k do not have the same class. Then we can replace $\hat{Q}$ by a diagonal $m\kappa^2 \times m\kappa^2$ matrix, $\tilde{Q}$, of full rank generically, provided we replace N_A by $\tilde{N}_A$ that is a $m\kappa^2 \times \kappa^3$ stacked matrix formed from the tensor product of each class component of $M^{\otimes_r^2}$ and M. From Figure 3.5, $N_B = N_A$ and can be replaced by $\tilde{N}_B = \tilde{N}_A = N$.

Since $\tilde{Q}$ is diagonal with positive entries, then

$$\text{Flat}_{A|B}(P) = N^T (\tilde{Q})^{1/2} (\tilde{Q})^{1/2} N$$

Let $\Lambda = (\tilde{Q})^{1/2}N$, so

$$\text{Flat}_{A|B}(P) = \Lambda^T\Lambda.$$

By the singular value decomposition, we see

$$\text{rank}(\Lambda^T\Lambda) = \text{rank}(\Lambda) = \text{rank}(N).$$

A `PARI/GP` calculation presented as Proposition 21 shows that $\text{rank}(N) > m\kappa$ generically. Thus generically

$$\text{rank}(\text{Flat}_{A|B}(P)) > m\kappa.$$

Now suppose that $|A'| = |B''| = 2$ and $|A''| = |B'| = 1$, as shown in Figure 3.6. The argument for the first case does not work for this tree because the tensor products for N_A and N_B are different, as the tensor products are taken in different orders. However a more complicated `PARI/GP` calculation, presented as Proposition 22 shows that $\text{Flat}_{A|B}(P)$ generically has rank greater than $m\kappa$.

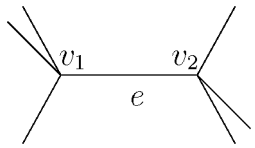

Figure 3.6: A tree such as in Figure 3.4 with $|A'| = |B''| = 2$ and $|A''| = |B'| = 1$.

For the general case of $|A|, |B| \geq 3$, take $\hat{A}$ to be a 3-element subset of A with at least one element from A' and from A'' according to Figure 3.4 and similarly take $\hat{B}$ to be a 3-element subset of B with at least one element from B' and from B''. Since the row indices of $\text{Flat}_{A|B}(P)$ depend on the states at the taxa in A and the column indices depend on the states at the taxa in B, marginalizing over all possible states for the taxa in A which are not in $\hat{A}$ and same for B, gives us $\text{Flat}_{\hat{A}|\hat{B}}(\hat{P})$. Then there are some appropriate matrices, J_1, J_2 which perform this marginalization on $\text{Flat}_{A|B}(P)$,

$$J_1 \,\text{Flat}_{A|B}(P) J_2 = \text{Flat}_{\hat{A}|\hat{B}}(\hat{P}).$$

Since $\text{Flat}_{\hat{A}|\hat{B}}(\hat{P})$ has rank greater than $m\kappa$ and $\text{Flat}_{A|B}(P)$ has rank greater than or equal to $\text{Flat}_{\hat{A}|\hat{B}}(\hat{P})$, then $\text{Flat}_{A|B}(P)$ has rank greater than $m\kappa$. □

By Proposition 27, from a distribution P for generic parameters we can identify every edge in the tree for which there are at least three taxa on either side. Since we will need to identify a tripartition on the tree later, in the following proposition, we prove that Proposition 27 also helps us to find a tripartition on the tree.

Proposition 28. *If T has $n \geq 9$ taxa, then it has a tripartition we can detect from a distribution P using Proposition 27, for generic parameters.*

Proof. According to Lemma 4.8 in [5], every unrooted binary tree T with $n \geq 3$ has an internal vertex which induces a tripartition $A|B|C$ such that two of the three components contain at least $\lceil n/4 \rceil$ leaves of T. Then the third component has at least 1 leaf and at most $n - 1 - \lceil n/2 \rceil$ leaves. Note that for $n \geq 9$, $n - 1 - \lceil n/2 \rceil \geq 3$.

The two edges by which the components with at least $\lceil n/4 \rceil$ leaves join T can be detected by Proposition 27 since for $n \geq 9$, $\lceil n/4 \rceil \geq 3$. The edge to the third component can also be determined when it has greater than or equal to 3 leaves of T. We just need to argue about two cases when the number of taxa in this component is 1 or 2. The argument is illustrated for $n = 9$ in Figure 3.7.

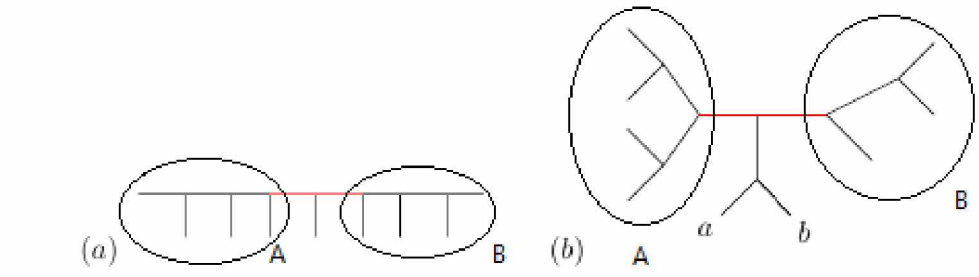

Figure 3.7: 9-taxa trees in which the number of taxa in the third component of a tripartition is 1 or 2.

If the third component has only one leaf, as in tree (a) of Figure 3.7, the two bipartitions $A \cup \{a\}|B$ and $A|B \cup \{a\}$ are identifiable by Proposition 27. These two imply the tripartition

for this tree is $A|B|\{a\}$. If the third component has two leaves as in tree (b) of Figure 3.7, the two bipartitions $A \cup \{a,b\}|B$ and $A|B \cup \{a,b\}$ are identifiable, but $A \cup \{a\}|B \cup \{b\}$ and $A \cup \{b\}|B \cup \{a\}$ are not displayed on T, and that is detected by Proposition 27. This implies the tripartition $A|B|\{a,b\}$ is on the tree. □

Now that we can identify a tripartition on a tree T such that two of three partitions have at least 3 taxa in them, the next step is to show we are able to apply Kruskal's theorem for one of these tripartitions. For a particular choice of parameters of the PM model, the Markov matrices describing the conditional probabilities of jointly observing the states in these partitions look like those in Proposition 18 which have full row rank, and the following lemma gives a precise proof for that.

Lemma 29. *For $\kappa = 20$, let R be the matrix of all 1's. Then for $m \leq 77$, $t_e = 1$, and $\ell \geq 3$, the row tensor ℓ^{th} power of a stacked Markov matrices under the PM model has full row rank for generic parameter choices of the $\boldsymbol{\pi}_i, r_i$.*

Proof. For a class i, we have $\kappa - 1$ and 1 independent parameters corresponding to π_i and r_i. Then with R and branch lengths fixed the dimension of parameter space for the m class PM model (not including class size frequencies) is

$$\hat{D} = m\big((\kappa - 1) + 1\big) = m\kappa.$$

Suppose U is a full-dimensional subset of our model's parameter space, $U \subseteq \mathbb{R}^{\hat{D}}$, as shown in Figure 3.8. We define a map ψ from U to an algebraic space of m-stacked Markov matrices of size $\kappa \times \kappa$ with rows adding to 1. This invertible analytic map takes $\boldsymbol{\pi}_i$'s, and $\{r_i\}$'s and maps them to a m-stacked Markov matrix as we compute Q_i's, exponentiate $r_i Q_i$'s, and stack them. Then $\psi : U \to \mathbb{R}^L \subseteq \mathbb{C}^L$ where $L = m\kappa(\kappa - 1)$.

Let V_0 be the Zariski closure of m-stacked Markov matrices of the form $M(a_1, \cdots, a_\kappa)$. This algebraic variety has dimension $m\kappa$. Note that by Lemma 16, $\psi(U) \subseteq V_0$ and has the same dimension, $\hat{D}$. Then $\psi(U)$ is a full-dimensional subset of $V_0 \cap \mathbb{R}^L$.

Now we consider the map that takes a m-stacked Markov matrix and maps it to its row tensor cube. Then let ϕ be the row tensor cube map on $m\kappa \times \kappa$ matrices, $\phi : \mathbb{R}^L \to \mathbb{R}^D$ where

$$D = m\kappa \times \binom{3+\kappa-1}{\kappa-1} = \frac{1}{6} m\kappa^2(\kappa+2)(\kappa+1),$$

with $\binom{3+\kappa-1}{\kappa-1}$ is the number of cubic monomials in κ variables. Then ϕ is a polynomial map and finite-to-one.

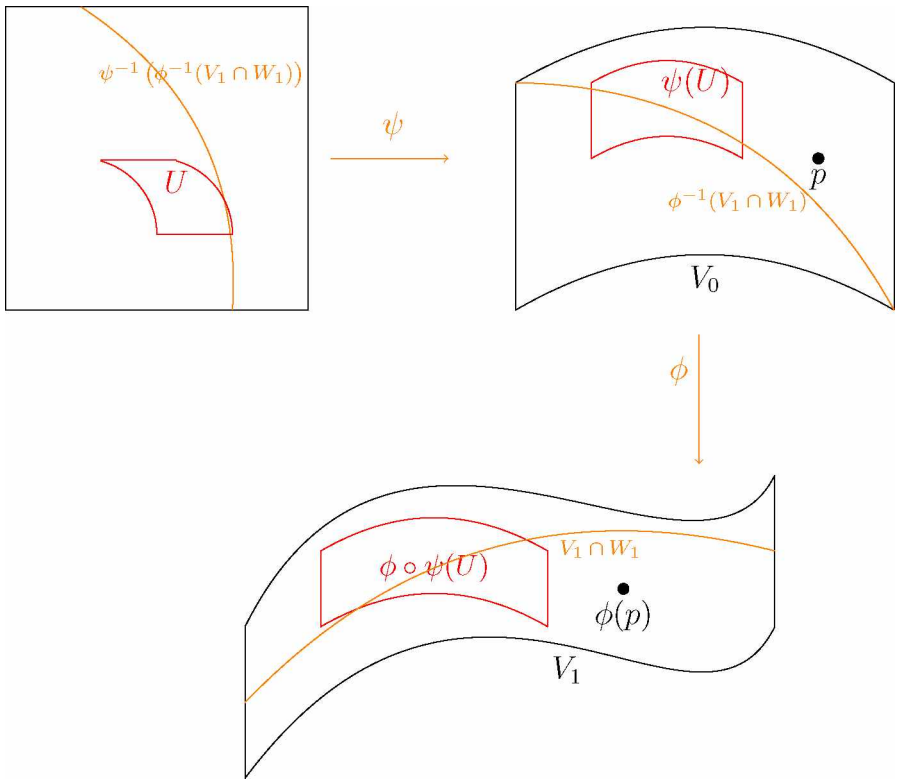

Figure 3.8: Schematic representation for Lemma 29 of the decomposition of the parametrization maps from stochastic space by ψ into an algebraic space of Markov matrices and then by ϕ to a probability distribution space. Here $V_1 \cap W_1$ is the set of the potential exceptional points where row tensor powers may not have full row rank.

The Zariski closure of the image of $\phi(V_0)$ is an algebraic variety $V_1 = \overline{\phi(V_0)}$. Then $\phi \circ \psi(U)$ is a patch in V_1 of dimension $\hat{D}$.

Proposition 18 shows that there exists a $p \in V_0$ with $\phi(p)$ having full row rank $m\kappa$. Then at least one of the $m\kappa \times m\kappa$ minors of $\phi(p)$ is nonzero. Call this minor, as a polynomial in the entries of $\phi(p)$, f. Let $W_1 = V(f)$ be the variety where f is zero. Since $V_1 \cap W_1 \subsetneq V_1$,

by Proposition 15 $V_1 \cap W_1$ has lower dimension than V_1. Since ϕ is finite-to-1, this implies $\phi^{-1}(V_1 \cap W_1)$ is a lower dimensional subvariety of $\phi^{-1}(V_1) = V_0$. Since ψ is $1-1$, this then implies $\psi^{-1}(\phi^{-1}(V_1 \cap W_1))$ is of dimension $< \hat{D}$. But any point $u \in U \setminus \psi^{-1}(\phi^{-1}(V_1 \cap W_1))$ is such that $f \circ \phi \circ \psi(u) \neq 0$, and thus $\phi \circ \psi(u)$ has full row rank. That is, generic points $u \in U$ are such that $\phi \circ \psi(u)$ has full row rank. □

Corollary 30. *Let $\kappa = 20$, $m < 77$, $t_e = 1$, $\ell \geq 3$. Then for generic $R, \boldsymbol{\pi}_i, r_i$, the row tensor ℓ^{th} power of stacked Markov matrices under the PM model has full row rank.*

Proof. The Zariski closure of the set of such parameterized matrices is an analytic variety. Since Lemma 29 shows there exist points on it of full rank, then Proposition 15, shows that generic points have full rank. □

We can similarly obtain from Proposition 20,

Corollary 31. *For generic $R, \boldsymbol{\pi}_i, r_i, \ell \geq 1$, the row tensor ℓ^{th} power of stacked Markov matrices under the PM model has Kruskal row rank ≥ 2.*

Now let T be an n-taxon binary tree on X. Picking any internal vertex of T, gives a tripartition of X that allows us to form 3 agglomerate observed variables under the PM model. Then we can attempt to apply Kruskal's theorem for the PM model. Suppose that P is a probability distribution in the PM model on tree T displaying tripartition $A|B|C$ of the leaves. Then one can give $m\kappa \times \kappa^{|A|}, m\kappa \times \kappa^{|B|}, m\kappa \times \kappa^{|C|}$ stochastic matrices M_A, M_B, M_C of conditional probabilities of states at the leaves in A, B, C given the state at v from the parameters on T. Then by Lemma 24, we know that

$$\text{Flat}_{A|B|C}(P) = [\boldsymbol{\pi}; M_A, M_B, M_C] = [\tilde{M}_A, M_B, M_C],$$

where $\tilde{M}_A = \text{diag}(\boldsymbol{\pi}) M_A$.

To use Kruskal's theorem on this flattening, we first show that Kruskal rank of the matrices are large enough, at least generically, that the theorem applies.

Lemma 32. *Let $\kappa = 20$ and $m \leq 77$. Let $\tilde{M}_A, M_B, M_C$ be those matrices described above. For generic numerical parameters of the PM model when $|A|, |B| \geq 3$, $|C| \geq 1$, $\tilde{M}_A$, M_B have full Kruskal row rank and M_C has Kruskal row rank equal to or greater than 2.*

Proof. Finding a single choice of parameters with this property is enough to get the result by similar reasoning as was used in Lemma 29. Let branch length on all internal edges be 0, so Markov matrices are I. Fix edge length for all terminal edges, (say, $t_e = 1$), and choose all entries of R to be 1. Pick $\boldsymbol{\pi}_i$'s entries to be small distinct positive numbers. Then T is a star tree, rooted at the central node, v.

For this choice of parameters, M_A, M_B, M_C are Markov matrices in the stacked form of Definition 4. Then by Corollary 30, for generic parameters M_A and so $\tilde{M}_A$, M_B have full row rank and so full Kruskal row rank when $|A|, |B| \geq 3$. Also by Corollary 31, M_C has Kruskal row rank equal to or greater than 2 when $|C| \geq 1$. □

We are now ready to prove our main result on the generic identifiability of numerical parameters of trees with a known tripartition.

Proposition 33. *Suppose T is a binary tree on X which displays a known tripartition $A|B|C$ with $|A|, |B| \geq 3$, $|C| \geq 1$, and $m \leq 77$. Then both T and the numerical parameters of the $PM(T, 20, m)$ model on T are generically identifiable.*

Proof. By Lemma 32, if a distribution P comes from generic parameters of the PM model on T, then

$$\text{Flat}_{A|B|C}(P) = [\tilde{M}_A, M_B, M_C],$$

where $\tilde{M}_A, M_B$ have full Kruskal row rank, M_C has Kruskal rank greater than 2. Thus equation 3.3 of Kruskal's Theorem is satisfied with $l = m\kappa$ and $[\tilde{M}_A, M_B, M_C]$ determines $\tilde{M}_A, M_B, M_C$ uniquely up to simultaneous permutation and scaling of the rows.

Also by factoring out row sums from the matrices, we can identify the root distribution matrix Π_r and M_A, M_B, M_C up to permutation of rows. If we pick the first diagonal entry

of Π_r and suppose that this row corresponds to class i and state ℓ, then the first row of M_A, M_B, M_C corresponds to the same class i and state ℓ. Kruskal's theorem does not give us order of rows based on the classes or the structure of subtrees, but we will find which rows of Markov matrices go together in the same class and put them together.

We know that $|A| \geq 3$. Consider first the case of $|A| = 3$, $A = \{a, b, c\}$. So we have the subtree of Figure 3.9 such that $\{x, y, z\} = \{a, b, c\}$.

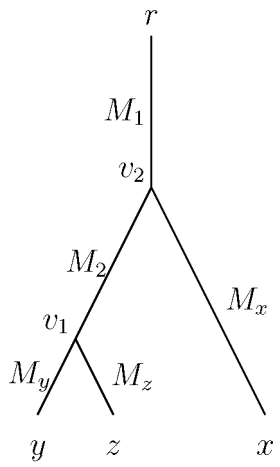

Figure 3.9: A subtree with 3 taxa and Markov matrices associated with the edges.

Then M_A is a $m\kappa \times \kappa^3$ matrix. Choose the ℓ^{th} row of M_A. It is a row vector with κ^3 entries, but we can reformulate this row as a 3-dimensional tensor, Γ, whose size is $\kappa \times \kappa \times \kappa$. This is the tensor giving a conditional distribution with i, j, k entries $P(a = i, b = j, c = k | r = \ell)$ when ℓ denotes both class and state. Now take v_1 as the root of the subtree for A shown in Figure 3.9. Then for unknown $p, M_x, M_y, M_z, M_1, M_2$ we can express the joint distribution of states at x, y, z, r as

$$
\begin{aligned}
P(x=i,y=j,z=k,r=\ell) &= \sum_{\alpha=1}^{\kappa}\sum_{\beta=1}^{\kappa} p_{v_1}(\beta)M_y(\beta,j)M_z(\beta,k)M_2(\beta,\alpha)M_1(\alpha,\ell)M_x(\alpha,i) \\
&= \sum_{\beta=1}^{\kappa} p_{v_1}(\beta)M_y(\beta,j)M_z(\beta,k)\left(\sum_{\alpha=1}^{\kappa} M_2(\beta,\alpha)M_1(\alpha,\ell)M_x(\alpha,i)\right) \\
&= \sum_{\beta=1}^{\kappa} p_{v_1}(\beta)M_y(\beta,j)M_z(\beta,k)\tilde{M}(\beta,i) \\
&= [\hat{p}_{v_1}\sigma^T;\sigma M_y,\sigma M_z,\sigma\hat{M}],
\end{aligned}
$$

where $\hat{M} = M_2\,\mathrm{diag}(M_1(:,\ell))M_x$. Let $\tilde{M}$ be $\hat{M}$ with rows normalized to add to 1, so for some D, $\tilde{M} = D\hat{M}$ and $\hat{p}_{v_1} = Dp_{v_1}$. Note that this joint distribution is just a rescaling of the conditional distribution given in the ℓ^{th} row of M_A with same ordering of the indices.

By applying Kruskal's theorem to each row of M_A reshaped into tensor, then we can decompose $P(x=i,y=j,z=k|r=\ell)$ for each ℓ with $1 \leq \ell \leq m\kappa$ into a 3-way product. Note that for each ℓ this deals only with a single class and the Markov matrices are then $\kappa \times \kappa$ and generically of full rank. So Kruskal's theorem gives the matrices $M_y, M_z, \hat{M}$ up to ordering of rows. Two of them, M_y, M_z will be dependent only on the class, but not the state of ℓ. So for different ℓ, we can find κ rows with the same (possibly permuted rows) version of M_y and M_z which give one class. In this way we can group the rows of M_A, M_B, M_C by class. Now taking those rows of M_A, M_B, M_C, and Π_r for one class and multiply them back together in a 3-way product gives a tensor for a single class GTR model. Both the tree and numerical parameters are identifiable for this model by Theorem 1.

For the case $|A| > 3$, this shows that by marginalization down to $|A| = 3$ as in Proposition 27 we can identify the subtrees and parameters for B, C. Then interchanging the roles of A and B gives the subtree and parameters for A. □

Combining Propositions 28 with Proposition 33 we have the main result.

Theorem 3. *Let T be a tree with at least 9 taxa. Then under the PM $(T, 20, m)$ model*

with $m \leq 77$, T and numerical parameters are identifiable from a probability distribution for generic parameters.

To complement our argument about the identifiability of model parameters, we indicate why the number of taxa of a tree satisfying in Proposition 33 should be greater than or equal to 9. In the assumption of Proposition 33 we have that T is a binary tree on X which displays a tripartition $A|B|C$ with $|A|, |B| \geq 3$, and $|C| \geq 1$. Then T needs to have at least 7 taxa. But when $|X| = 8$, there are 5 tree shapes as shown in Figure 3.10.

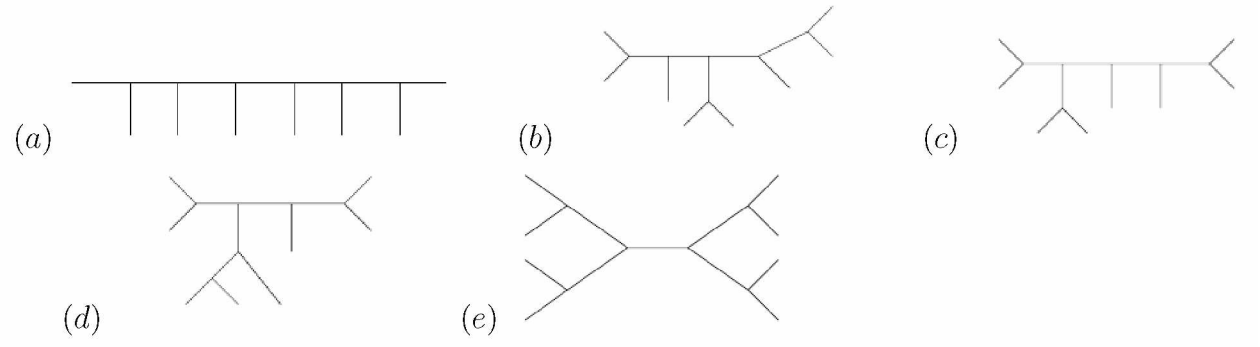

Figure 3.10: 8-taxon tree shapes

As it can be seen for tree (e) there is no tripartition $A|B|C$ with $|A|, |B| \geq 3$, and $|C| \geq 1$. Then Propositions 27 and 28 cannot be used to obtain a tripartition in order to apply Proposition 33. For the other 8-taxon trees shown, however, we can obtain identifiability.

3.7 Some other results

To further understand of the main result on identifiability of the PM model of this chapter, it is helpful to look at its relationship to other published results about parameters identifiability of phylogenetic models.

Allman and Rhodes [1] introduced a (λ, κ)-state general Markov model, $\mathcal{M}_{\lambda,\kappa}$, which is motivated by the covarion model of Tuffley and Steel [8]. In this model, internal nodes are allowed to have more states than leaves. Suppose T is a binary topological unrooted tree. Choose one of the internal nodes as a root, r, and call the tree T^r, with each edge directed from the root. Each leaf of the tree has an observed random variable with state

space $[\kappa] = \{1, 2, \ldots, \kappa\}$ and each internal vertex has an unobserved random variable with state space $[\lambda]$ such that $\lambda \geq \kappa$. Let π_r be a row vector with λ elements which represents the probability distribution of the states at the root with entries sum to 1. For each directed pendant edge, e, there is a $\lambda \times \kappa$ Markov matrix describing transition probabilities and a $\lambda \times \lambda$ matrix for each directed internal edge, e. Then Allman and Rhodes showed that for $\lambda < \kappa^2$ under an analytic (λ, κ)-state model on a 4-leaf tree, the tree topology is identifiable for generic parameters. In the context of the PM model, this gives the following.

Theorem 4. *For $m < \kappa - 1$ under the PM model, the 4-leaf species tree topology is identifiable from sequences.*

Thus their result for protein sequences with $\kappa = 20$ gives identifiability for at most 19 classes, which is insufficient to apply to the PM $(T, 20, 60)$ model used in biological applications.

Allman et al. [2] introduced another model named a mixture of coalescent mixtures. Their result at first appears to be surprisingly close to what have been proved in this thesis in that it shows identifiability of trees for very general many class mixtures. However, they need to assume the tree is ultrametric, which biological experience has shown is often violated. While they also allowed for a coalescent model, that is not relevant to this work, and is not necessary for their proof.

In the setting of this thesis their result gives the following.

Theorem 5. *Let $\kappa \geq 2$, $m \in \mathbb{N}$, and T be a binary ultrametric tree. Then under the PM (T, κ, m) model, the tree topology is identifiable from the expected log-det distance for pairs of taxa.*

While the proof of this theorem is much less complicated than the approach given in this thesis, its restriction to ultrametric trees and failure to identify numerical parameters seems essential to the approach.

References

[1] Allman, E. S., and J. A. Rhodes (2006), The identifiability of tree topology for phylogenetic models, including covarion and mixture models, *Journal of Computational Biology*, *13*, 1101–1113, doi:10.1007/s00285-010-0355-7.

[2] Allman, E. S., C. Long, and J. A. Rhodes (2018), Species tree inference from genomic sequences using the log-det distance, *Journal of Mathematical Biology*, *62*, 833–862, doi: 10.1007/s00285-010-0355-7.

[3] Kruskal, B. J. (1977), Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, *Linear Algebra and its Applications*, *18*(2), 95 – 138, doi:https://doi.org/10.1016/0024-3795(77)90069-6.

[4] Quang, L. S., O. Gascuel, and N. Lartillot (2008), Empirical profile mixture models for phylogenetic reconstruction, *Bioinformatics*, *24*(20), 2317–2323, doi:10.1093/ bioinformatics/btn445.

[5] Rhodes, J. A., and S. Sullivant (2012), Identifiability of large phylogenetic mixture models, *Bulletin of Mathematical Biology*, *74*, 212–231, doi:10.1007/s11538-011-9672-2.

[6] Semple, C., and M. Steel (2003), *Oxford lecture series in mathematics and its applications*, vol. 24.

[7] Tavaré, S. (1986), Some probabilistic and statistical problems on the analysis of DNA sequences, *Lectures on Mathematics in the Life Sciences*, *17*, 57–86.

[8] Tuffley, C., and M. Steel (1998), Modeling the covarion hypothesis of nucleotide substitution., *Mathematical Biosciences*, *147*, 63–91.

Chapter 4: Conclusion and Future Work

In this work we introduced a new method of inferring metric species trees from topological gene trees under the Multispecies Coalescent (MSC) model. This algorithm provides a statistically consistent estimator of a species tree by defining new intertaxon distances that can be calculated from the weights of quartets for all subsets of four taxa. By using these distances, this method produces a metric species tree that exactly fit the same tree topology, but with rescaled edge weights by certain factors.

A natural problem is to extend this method from trees to phylogenetic networks. Networks are different from phylogenetic trees since they have hybrid nodes (nodes with two parents) instead of only tree nodes (nodes with only one parent). The quartet distance [2] which gives only topological information of species trees has been generalized to the level-1 network setting in the NANUQ method [3]. NANUQ takes as input a set of gene trees and produces an unrooted network with certain properties. NANUQ is a recent product of the biomathematics research group in the Department of Mathematics and Statistics at UAF but there is much potential to extend it. Extending the metric quartet distance seeks to improve on this method to provide metric information for networks as well.

Also from this project, it is clear that gene tree error is a significant contributor to the lack of accuracy of inferred species trees. Then focusing on the errors in sampling gene trees from MSC can be future work. Better ways are needed to quantify how much of gene tree variation could be due to the MSC and how much is simply error can be investigated.

We also proved that both numerical and non-numerical parameters of a model used for inferring a species tree from protein sequences are identifiable. Thus, another interesting future question is to explore the identifiability problem for models used for inferring the species network. There are some results for identifying the topology or topological features of the species network, but not all parameters.

References

[1] Wang, Huai Chun and Minh, Bui Quang and Susko, Edward and Roger, Andrew J. *Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation.* Systematic Biology, 2018.

[2] E.S. Allman and H. Baños and J.A. Rhodes *NANUQ: A method for inferring species networks from gene trees under the coalescent model.* Algorithms Mol. Biol., 2019.

[3] J.A. Rhodes. *Topological metrizations of trees, and new quartet methods of tree inference.* IEEE/ACM Trans. Comput. Biol. Bioinform.,2019

PARI/GP functions needed for computations

```
\\ Make an nxn matrix of the form M(a1, ...., an) where v holds the ai
matAlgRandom(n,x)={        \\ n=the length of vector v , x= a random number
my(v,A,B,S);
v=vector(n);
A=matrix(n,n);       \\ a square matrix whose each row is the same as v
S=matrix(n,n);       \\ sum of A and the diagonal matrix B
\\ filling the entries of v with random numbers between 0 and x
for(i=1,n, v[i]=random([1,x])); c=vecsum(v);
for(i=1,n, for(j=1,n, A[i,j]=v[j] ));
B=(1-c)*matid(n);   \\ a diagonal matrix with 1-c on diagonal
S=A+B;
return(S);
}
matAlg(n,v)={
my(M,s);
s=vecsum(v);
M=matrix(n,n)
for(i=1,n, for(j=1,n, M[i,j]=v[j] ));
for(i=1,n, M[i,i]=1-s+M[i,i]);
return(M);
}
\\ Row Kronecker
matRowKron(M,N)={
my(m=#M[,1]);
my(mCols=#M[1,]);
my(nCols=#N[1,]);
my(numCols=mCols*nCols);
\\ for returning the row tensor product
my(L=matrix(m,numCols));
for(i=1,m, for(j=1,mCols, for(k=1,nCols,L[i,(j-1)*mCols+k]=M[i,j]*N[i,k] )));
return(L);
}
matKron(M,N)={
my(mRows=#M[,1]);
my(nRows=#N[,1]);
my(mCols=#M[1,]);
my(nCols=#N[1,]);
my(numCols=mCols*nCols);
\\ for returning the row tensor product
```

```
my(L=matrix(mRows*nRows,mCols*nCols));
for(i=1,mRows, for(ii=1,mCols,
for(j=1,nRows,
for(jj=1,nCols,L[(i-1)*nRows+j,(ii-1)*nCols+jj]=M[i,ii]*N[j,jj] ))));
return(L);
}
\\ Reduced Row Tensor product power 2
matRedRow2(M)={
my(m=#M[,1]);
my(n=#M[1,]);
nL=binomial(n+1,n-1);
my(L=matrix(m,nL));
for(i=1,m, l=1;  for(j=1,n,  for(k=j,n, L[i,l]=M[i,j]*M[i,k]; l++;)));
return(L);
}
\\ Reduced Row Tensor product power 3
matRedRow3(M)={
my(m=#M[,1],n=#M[1,],L);
nL=binomial(n+2,n-1);
L=matrix(m,nL);
for(i=1,m, l=1; for(j=1,n,
for(k=j,n,
for(s=k,n, L[i,l]=M[i,j]*M[i,k]*M[i,s]; l++;) )));
return(L);
}
\\ Compute some ranks
rankComputations2(AA)={
BB=matRedRow2(AA);
print("\n The size of the matrix is");
print(matsize(BB));
BBrank=matrank(BB);
print("The rank is ");
print(BBrank);
}
\\ Compute the rank
rankComputations3(AA)={
BB=matRedRow3(AA);
print("\n The size of the matrix is");
print(matsize(BB));
BBrank=matrank(BB);
print("The rank is ");
print(BBrank);
}
```

PARI/GP codes needed for Props 18 and 21

```
\\ Compute rank of matrices needed for Props 18 and 21
\\ \r "/Users/eallman/Dropbox/samaneh research/snForm/pari/startup.gp";
\\ default(parisize,1G);
default(parisize,4G);
default(parisizemax,8G);
/* Using log files  */
default(logfile,"/Users/eallman/Desktop/matRank_gp.log");  \\ turn logfile on
default(log,1);
\\ start timer
#
kappa=20;
numClasses=4;
\\ for Proposition 18
print("\n\n****** reduced row THIRD tensor powers *****");
A=[;];
\\for (j=1,numStacked,stackNumber=j;B=matAlgRandom(kappa,30);
A=matconcat([A;B]);rankComputations3(A););
for (j=1,78,stackNumber=j;B=matAlgRandom(kappa,30);
A=matconcat([A;B]);rankComputations3(A););
\\ end timer
#
\\ for Proposion 21
\\ start timer
#
print("\n **** kappa is ",kappa," and number of Classes is ",numClasses,"\n");
A=[;];
for (m=1,numClasses,classNum=m;B=matAlgRandom(kappa,70);
C=matRedRow2(B);B=matKron(C,B);
A=matconcat([A;B]);printp("Number of classes is ",m);
printp(" Size should be [",m*kappa^2, ", ", binomial(kappa+1,kappa-1)*kappa,"]");
printp(" Size of A is ",matsize(A));r=matrank(A);
print(" Rank of A is ",r," and m*kappa is ",m*kappa);
if (r>m*kappa,printp("Check.\n"),printp(" ** Rank deficient ***\n")););
\\ end timer
#
default(log,0);  \\ turn logfile off
\\ \q
```